Aalto University
School of Business

THE ROLE OF BENCHMARKING FAILURES IN MUTUAL FUNDS' DECREASING ECONOMIES OF SCALE

Evidence from the U.S. actively managed equity mutual funds

Master's Thesis
Niko Pronin, Tiia Sinisalo
Aalto University School of Business
Finance
Spring 2021

Aalto University School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Authors Niko Pronin, Tiia Sinisalo

Title of thesis The role of benchmarking failures in mutual funds' decreasing economies of scale

Degree Master of Science in Economics and Business Administration

Degree programme Finance

Thesis advisor Peter Nyberg

Year of approval 2021 **Number of pages** 58 **Language** English

Abstract

Recent mutual fund literature describes irrational fund flow patterns apparently due to investors' tendency to reward fund managers with excess fund flows for higher materialised factor-related returns. Most interestingly for this study, Song (2020) links these fund flow patterns with the decreasing economies of scale hypothesis in mutual funds by proposing that the excess fund flows have significant predictive power over the funds' future benchmark-adjusted performance. The main contribution of our study is to figure out to which extent the documented inverse relationship between mutual funds' prior excess fund flows and their subsequent future benchmark-adjusted performance is a consequence of fund managers' inability to replicate high factor-related returns attributable to luck in future, while investors are incapable of realising this via their benchmarking attempts.

We perform the study with actively managed equity mutual funds obtained from The Center of Research in Security Prices (CRSP) Survivor-Bias-Free Mutual Fund Database between 1992–2019. Using Carhart's (1997) methodology, we construct share class portfolios according to the estimated factor-related returns, skill, and size of the mutual funds' share classes. We continue to estimate future abnormal returns of these share class portfolios to infer if any performance differences occur between the portfolios ex-post the factor-return estimation periods. To test the factor-returns' non-replicability hypothesis we apply a set of fixed effects panel regressions and time series regression analyses, while in the robustness section we utilise Fama Macbeth (1973) regression analyses.

Throughout the study, we find evidence that aligns with the Benchmarking Misevaluation Hypothesis (BMH), a theory that we propose and study as an alternative explanation for the inverse relationship between prior factor-related returns and future benchmark-adjusted performance. We find that mutual funds' share classes with the highest (top) prior factor-related returns underperform their benchmarks significantly in the future by up to 158 bps while share classes with the least prior factor-related returns do not, despite that the top share classes seem to neither attract significant excess fund flows relative to their size nor to become the largest measured in earmarked total net assets. Also, it appears that the most prominent BMH-associated return patterns prevail among smaller mutual fund share classes, suggesting that their sub-optimally diversified holdings push these share classes to the top third measured in factor-related returns when the risks realise favourably. Moreover, the sophistication that prevails among the funds' investor bases seems to impact to which extent the share classes exhibit BMH-associated return patterns, suggesting that investors' lack of skill in benchmarking is linked with the phenomenon. Finally, the changes in the return patterns appear to react to market conditions as we expect, although the significance of recessions in the BMH-associated patterns are more substantial compared to what we hypothesise.

Nonetheless, the evidence remains inconclusive. Despite repeatedly favourable patterns, we find insignificant evidence for the difference in performance between share classes with the highest and the lowest prior factor-related returns in the future. Moreover, the estimates of underperformance are not substantial enough to explain the differences of 300–400bps that Song (2020) reports in his paper. Thus, we conclude that although the BMH seems to prevail in mutual funds to some extent, it lacks robustness to disposition earlier conclusions that link the inverse relationship between prior factor-related returns and future benchmark-adjusted performance with the decreasing economies of scale hypothesis.

Keywords mutual funds, decreasing economies of scale, benchmarking misevaluation hypothesis, fund flows

Aalto-yliopisto
Kauppakorkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Tekijät Niko Pronin, Tiia Sinisalo

Työn nimi Vertailuanalyysivirheiden osuus sijoitusrahastojen laskevissa mittakaavaeduissa

Tutkinto Kauppatieteiden maisterintutkinto

Koulutusohjelma Rahoitus

Työn ohjaaja Peter Nyberg

Hyväksymisvuosi 2021 **Sivumäärä** 58 **Kieli** Englanti

Tiivistelmä

Viimeaikainen sijoitusrahastokirjallisuus on käsitellyt runsaasti varallisuusvirtoja, jotka aiheutuvat sijoittajien taipumuksesta palkita rahastonhoitajia korkeista realisoituneista riskifaktoriliitännäisistä tuotoista. Tämän tutkimuksen kannalta mielenkiintoisimpia edeltäjiä on Songin (2020) havainto, jolla hän linkittää epärationaaliset varallisuusvirrat sijoitusrahastojen laskeviin mittakaavaetuihin ehdottamalla, että edeltävillä ylisuurilla varallisuusvirroilla on merkittävää ennustevoimaa tulevien ajanjaksojen riskikorjattuihin tuottoihin. Tämän tutkimuksen pääkontribuutio on selvittää, kuinka olennaisesti aiemmin kuvattu käänteinen riippuvuussuhde aikaisempien ylisuurien varallisuusvirtojen ja tulevien riskikorjattujen tuottojen välillä on seurausta rahastonhoitajien kyvyttömyydestä toistaa tulevaisuudessa aiempia onnekkaasti korkeita faktoriliitännäisiä tuottoja, mitä rahastojen sijoittajat eivät puolestaan havaitse vertailuanalyysiyrityksistään huolimatta.

Suoritamme tutkimuksen aktiivisesti hallinnoiduilla osakerahastoilla, jotka saamme The Center of Research in Security Prices (CRSP) selviytymisharhavapaasta sijoitusrahastotietokannasta 1992–2019 väliltä. Carhartin (1997) metodologiaa hyödyntämällä muodostamme osakesarjaportfolioita arvioitujen faktorituottojen, taidon ja korvamerkittyjen varojen mukaan ja vertailemme niiden riskikorjattuja tuottoja tulevina ajanjaksoina tarkastellaksemme, onko merkittäviä suoriutumiseroja havaittavissa näiden portfolioiden välillä. Tarkastellaksemme tarkemmin korkeiden faktorituottojen toistumattomuushypoteesia toteutamme joukon kiinteiden vaikutusten paneeliregressioanalyysejä, kun taas tulosten vahvistamiseksi hyödynnämme Fama Macbeth (1973) regressioanalyysejä.

Löydämme läpi tutkimuksen näyttöä, joka tukee vertailuanalyysivirheiden hypoteesia (BMH), eli nimeämäämme teoriaa, jota tutkimme vaihtoehtoisena selityksenä laskevien mittakaavaetujen tuottosarjoille. Kun tarkastelemme sijoitusrahastoja yksittäisten osakesarjojen tasolla, korkeimmilla viimeaikaisilla faktoriliitännäisillä tuotoilla varustetut osakesarjat alisuoriutuvat riskikorjatusti 158 bps tulevina ajanjaksoina, vaikka kyseiset osakesarjat eivät houkuttele varoihinsa nähden ylisuuria varallisuusvirtoja tai kasva verrokkejaan suuremmiksi. Lisäksi merkittävimmät BMH-liitännäiset tuottosarjat ovat yleisempiä pienimpien osakesarjojen joukossa, mikä johtunee niiden heikommin hajautetuista omistuksista, jolloin näillä rahastoilla on aina korkeimmat faktorituotot riskien realisoituessa toivotusti. Edellä mainittujen havaintojen jatkoksi huomaamme, että sijoittajakannan sivistyksellä vaikuttaisi olevan merkitystä BMH-liitännäisten vaikutusten suuruuteen, mikä ilmentää sitä, että rahastosijoittajien vertailuanalyysitaidoilla on osansa ilmiössä. Lopuksi toteamme, että edellä mainitut tuottosarjat vaikuttavat noudattavan odottamaamme dynamiikkaa eri markkinatilanteissa, joskin taantuma-aikojen merkitys on odottamaamme suurempi.

Edellä mainituista lupaavista tuloksista huolimatta tämän tutkielman tulokset eivät ole ratkaisevia. Tämä johtuu siitä, ettemme löydä tilastollisesti merkitsevää eroa tulevaisuuden riskikorjatuille tuottotasoille vertaillessamme osakesarjoja, jotka ovat kuuluneet äskettäin joko korkeimpien tai matalimpien faktoriliitännäisten tuotollisten osakesarjojen joukkoon. Lisäksi arvioimamme suoriutumisero ei ole suuruusluokaltakaan tarpeeksi suuri selittämään 300–400 bps:n eroja, joita Song (2020) raportoi tutkimuksessaan. Tämän vuoksi päällimmäisin johtopäätöksemme on, että vaikka löydämme todisteita vertailuanalyysivirheiden hypoteesin läsnäolosta sijoitusrahastoissa, hypoteesi yksin ei kykene kumoamaan aikaisempia johtopäätöksiä, jotka liittävät faktoriliitännäiset tuotot ja ylisuuret varallisuusvirrat sijoitusrahastojen laskevien mittakaavaetujen hypoteesiin.

Avainsanat sijoitusrahastot, laskevat mittakaavaedut, vertailuanalyysivirheiden hypoteesi, varallisuusvirrat

Table of Contents

List of Tables

List of Figures

Key abbreviations

AUM = Assets Under Management

BMH = The Benchmarking Misevaluation Hypothesis

CAPM = Capital Asset Pricing Model

CRSP = The Center of Research in Security Prices

ETF = Exchange-Traded Fund

FED = The Federal Reserve System

FFC = Carhart (1997) benchmarking model

FRRs = Factor-Related Returns

HML = "High-Minus-Low" value factor

LHS = Left Hand Side

NAV = Net Asset Value

QE = Quantitative Easing

RHS = Right Hand Side

SMB = "Small-Minus-Big" size factor

TNA = Total Net Assets

UMD = "Up-Minus-Down" cross-sectional momentum factor

1 Introduction

"It's hard to detect good luck – it looks so much like something you've earned."
– Frank A. Clark

Mutual funds are an extensively researched area in the field of finance for vindicated reasons. Not only do they hold a significant figure in financial assets and thus are a material portion of the financial sector, but they also offer an extensive set of data reporting their holdings and historical performance. Therefore, mutual funds offer a broad set of research opportunities.

However, under the assumption of perfect capital markets, not much should be of great interest from the perspective of research. In market equilibrium, Berk & Green (2004) propose that mutual funds with the most skilled managers and superior performance attract the most fund flows. As the demand for the services of these fund managers increases, the managers raise their respective management fees, ultimately offering investors net returns that align broadly with the risk-adjusted benchmarks. Surprisingly, despite the large size of the sector and a versatile set of market participants, the previously proposed equilibrium is far from the empirical reality.

Market imperfections are persistent in the mutual fund industry. These arise from numerous sources, including agency costs due to information asymmetry between fund managers and investors, and the overall sophistication of the market participants in the mutual fund sector. For instance, approximately 89% of the mutual fund assets are owned by households alone in the United States in 2019[1], being naturally a considerable source of the imperfections that prevail in the mutual fund industry, such as return chasing, mental accounting, and other biases.

One of the most recent developments in mutual fund literature during the past decade is the observation that investors seem to reward fund managers with excess fund flows for regular factor-related returns (FRRs). The observation implies that mutual fund investors are generally incapable of differentiating return components into managers' skill, luck, and common factor exposures that earn risk premiums. One of the most recently documented consequences of the imperfect fund flows is its potential link to the mutual funds' decreasing economies of scale

[1] Investment Company Institute, 2019.

hypothesis that has been documented consistently in the prior literature. Most recently, Song (2020) argues that mutual funds with the highest FRRs attract excess fund flows compared to what is justified by the exhibited level of skill in the performance. Consequently, these funds grow beyond their respective fund managers' active management skill, leading these funds to underperform their relevant benchmarks in the future. Song (2020) further suggests that a significant portion of the future underperformance is attributable to increased aggregate trading costs, including liquidity demand, bid-ask spreads, limited capacity of trading strategies, as well as decreasing marginal returns on newly invested assets.

We hypothesise that a significant portion of the return patterns that Song (2020) reports can be attributable to fund managers' lucky high FRRs and investors' benchmarking mistakes instead of the decreasing economies of scale in mutual funds. We refer to the idea as *The Benchmarking Misevaluation Hypothesis* (BMH). The logic behind the hypothesis is that, on aggregate, those mutual funds that report the highest estimated FRRs over a certain period of time have been systematically luckier funds than those mutual funds with lower FRRs over the same period of time. However, mutual fund investors have challenges in recognising to what extent the recent performance of the fund managers is attributable to skill, luck, and common factor exposures that are supposed to earn risk premiums[2]. Since mutual fund investors seem to invest in funds with the best recent[3] track record (Berk & Green, 2004), the investors settle for investing in recently lucky mutual funds, i.e., funds that are less likely to replicate their performance going forward[4]. Hence, we would observe similar patterns as those in Song's (2020) paper, although the source of the return patterns are far from the decreasing economies of scale.

The phenomenon can be illustrated with people who claim to be exceptionally good at tossing a coin, having flipped four heads in a row. For the sake of illustration, we form a group of people who have tossed four heads in a row. We label them as skilled coin flippers, although we are not knowledgeable of their actual level of skill. We only have a backward-looking realisation of some probability distribution as a reference point, which is their historical track record. Say we have now another group of people who bet if the fifth toss of the coin flippers will be heads or tails. If the gamblers interpret wrongfully that the prior results are attributable

---

[2] See, e.g., Siegel, Waring & Scanlan, 2009; Berk & van Binsberg, 2016; Barber, Huang & Odean, 2016; Chakraborty, Kumar, Muhlhofer & Sastry, 2018.
[3] Further see Ben-David, Li, Rossi & Song, 2018.
[4] See section *5.2 Prior FRRs and future abnormal returns – lack of diversification and luck?*

to skill, they are inclined to favour the heads over tails systematically. As most coin flippers have achieved the prior results by other factors than skill, they are less capable of repeating the performance in the future, causing most gamblers to encounter losses. We argue that similar dynamics prevail in the mutual fund industry.

We hypothesise that the high FRRs are often attributable to sub-optimally diversified holdings of mutual fund share classes with smaller earmarked total net assets (TNA) (Pollet & Wilson, 2008), which enables luck to play a significant role in determining which mutual funds eventually belong to the group with the highest FRRs over their respective estimation periods. The main contribution of this thesis is to answer the main research question: "To what extent the inverse relationship between prior FRRs and future benchmark-adjusted performance is attributable to fund managers' inability to replicate prior lucky performance in the future, while investors are incapable of observing this?" The research idea of this thesis stems from that Song (2020) does not test whether his conclusions are robust on share class level, nor if the phenomenon behaves differently among mutual funds with varying level of financial sophistication in their respective investor bases. Thus, our thesis contributes to the existing literature by either strengthening the earlier findings of Song (2020) and other relevant prior publications or by shedding light on a new perspective that requires further attention in future research work.

This master's thesis is an extension of the bachelor's thesis of Pronin (2020), titled as *Think you can choose your asset manager? Think again – mutual fund investors' inability to distinguish fund managers' skill*. We perform our study in two steps. First, we show that treating each mutual fund share class as a single mutual fund, funds with the highest prior FRRs appear to attract significant excess fund flows when controlling for the exhibited skill of the funds, but only when measuring the average quarterly net fund flows in dollar value. However, these share classes neither attract any significant excess fund flows when normalising the magnitude of the flows to the size of the funds at the start of the observation periods nor grow to become the largest mutual funds among their respective peer groups, unlike Song (2020) reports in his study on aggregated mutual fund entities. These findings are essential in studying the relationship between prior FRRs and future benchmark-adjusted performance from the perspective of the BMH, as we already provide some evidence against the proposition of Song (2020).

Second, we study the relationship between prior FRRs and future risk-adjusted-performance. In addition to the complete sample analysis, we perform similar tests during varying levels of idiosyncratic volatility and prevailing financial sophistication in the investor bases. The findings are twofold. On the one hand, we find significant and consistent evidence of return patterns that align with the BMH. We find that regardless of the investor base sophistication and rational fund flow patterns, funds' share classes with the highest prior FRRs underperform their respective benchmarks significantly in the future, while funds with the least prior FRRs do not. On the other hand, we fail to find consistent evidence that the performance difference between high and low prior FRR share classes is significant, suggesting that the patterns are not robust enough to disposition earlier conclusions regarding the decreasing economies of scale hypothesis. Furthermore, the estimated order of magnitude in underperformance is not large enough to explain the performance difference of 300–400bps that Song (2020) reports in his paper. On top of that, we even find evidence of decreasing economies of scale hypothesis when we perform a set of robustness checks for the main findings of this thesis.

The findings are interesting from three perspectives. First, we find that luck seems to play only a limited role in the return patterns labelled as decreasing economies of scale. Second, we find evidence that fund flow dynamics are somewhat more rational in the mutual fund industry compared to some of the conclusions presented in prior literature. Third, we show that the impact of sub-optimally diversified holdings and luck in prior return realisations can nevertheless explain a portion of the widely documented underperformance of mutual funds measured in net risk-adjusted returns, emphasising the importance of that facet in future fund performance studies going forward.

2 Prior literature

2.1 Decreasing economies of scale in mutual funds

Decreasing economies of scale in mutual funds refers to a phenomenon where benchmark-adjusted returns of the funds delivered for their investors fall as the funds grow larger measured in total AUM. Berk & Green (2004) provide a pioneering explanation for the phenomenon in their theoretical framework for mutual fund flows under the assumption of rational markets. The key assumption of their model is that securities market is inefficient, providing abnormal return opportunities for extremely skilled fund managers, but the fund markets work efficiently, implying that fund investors tend to not enjoy abnormal returns regardless of the skill of the fund managers.

According to the model of Berk & Green (2004) , investors chase historical returns as a consequence of Bayesian inference of the fund managers' skill, concentrating fund flows in the most successful funds according to the most recent[5] performance figures. The concentration of fund flows, together with positive returns for the assets, grow these funds. As the demand increases for the services of the skilled managers, while these managers also command a higher share of the total mutual fund sector measured in TNA outstanding, their respective negotiation power over fund investors increases, leading fund managers to demand higher fees and compensation for their services. As Berk & Green (2004) assume efficient markets for fund investments, they propose that at an equilibrium state, management fees increase to the point that no abnormal returns are delivered for the fund investors. Rather, the ultimate beneficiaries of fund managers' exceptional skill are the skilled fund managers themselves.

Series of empirical studies induce from the preliminary proposal of Berk & Green (2004), supporting the hypothesis of decreasing economies of scale. Around the same time as Berk & Green (2004) published their model, Chen, Harrison, Huang & Kubik (2004) find that fund returns decline with fund size even after accounting for various performance benchmarks. However, unlike how Berk & Green (2004) suggest, Chen, Harrison, Huang & Kubik (2004) find that not only net returns fall but the gross performance of the funds also drops significantly.

[5] Further, see Ben-David, Li, Rossi & Song, 2018.

Zhu (2018) further finds a significantly negative impact that fund size has on its future alpha, indicating that the two are not independent measures.

Findings of Lou (2012) and Song (2020) are an interesting break through in potentially understanding the root cause of the decreasing economies of scales in the context of mutual funds. Both authors propose that rather than the fund size per se, it is how these mutual funds attract their flows that explains the future expected performance. In his paper, Song (2020) argues that instead of the AUM of the mutual funds, it is how much excess fund flows the fund managers attract attributable to biased reasons that impacts the future fund performance measured both on an absolute and on risk-adjusted basis. Song (2020) shows that those flows attracted by high prior FRRs predict significantly fund underperformance compared to its relevant benchmarks, while justified flows do not. But instead of remarking explicitly that the patterns are potentially a consequence of fund managers' inability to repeat prior lucky performance and that investors are incapable of realising that, he proposes that the drop in expected benchmark-adjusted returns is attributable to increasing trading costs among funds that have grown beyond the active management skill of their respective fund managers.

Mutual fund trading costs are relatively little researched area despite they are consistently being proposed to be the main explanation for the decreasing economies of scale. The reason behind it is two-fold. First, mutual funds are not obligated to report their trading costs in a similar manner as they are obligated to report management fees. Second, the real mutual fund trading costs are difficult to observe, as they consist of both direct and indirect portions of the total trading costs. The direct portion covers trading-related expenses, such as broker fees, while indirect costs arise from market liquidity-related issues, including price impact of larger trades, bid-ask spreads, *et cetera*.

The pioneering work in the field comes from Edelen, Evans, and Kadlec (2009, 2013). Their papers examine the role of trading costs as a source of diseconomies of scale by estimating annual trading costs for a sample of equity funds. The main findings are that trading costs have an increasingly disadvantageous impact on mutual fund performance as the relative trade size of the funds increases, being evidence in favour of the price impact theory of trading costs. They further find that trading costs are comparable in magnitude to the expense ratios, while they have higher cross-sectional variation that is directly related to the trade size of the funds, implying that some of the managers are more competent in managing larger AUM figures than the others.

Similar findings are made by Busse, Chordia, Jiang & Tang (2017) with their algorithm-based approach. They researched trading costs of the U.S. actively managed equity mutual funds, finding that the average figure is approximately 75bps per year, being persistent over time, and negatively related to fund performance. They further estimate that larger trades in smaller stocks and low-priced stocks incur higher transaction costs, further supporting the liquidity demand hypothesis in decreasing economies of scale in the context of mutual funds. However, counterintuitively they find that larger funds have lower trading costs than smaller funds despite their larger trade sizes. They find, however, that the larger funds endogenously hold and trade bigger and more liquid stocks, and trade less frequently, explaining the unanticipated relationship between mutual funds' trading costs and the size of the funds' AUMs.

The popularity of trading costs explaining the decreasing economies of scale comes as no surprise. Trading costs are a natural market imperfection inducing potentially numerous interesting anomalies, while the requirement of estimating the trading costs gives artistic liberty for academics in choosing appropriate methodology. Despite it is highly probable that trading costs explain a significant portion of the variation in benchmark-adjusted returns in mutual fund sector, the majority of the prior literature overlooks the impact that investors' benchmarking malfunctions might have on the decreasing economies of scale patterns in mutual funds.

An interesting detail in empirical work of mutual funds suggests that unsophisticated benchmarking could be driving the return patterns labelled as evidence of the decreasing economies of scale in mutual funds. While decreasing economies of scale is a widely documented phenomenon, Pollet & Wilson (2008) and Yan (2008) find that asset growth has little effect on the behaviour of typical mutual fund managers, although they find some evidence that a fraction of these funds diversifies their portfolios in response to growth. Even though the observation supports the theory of liquidity and price impact as a source of decreasing economies of scale, the observation also emphasises the relevance of studying the impact of the BMH. The fact that mutual funds diversify their holdings as response to growth suggests that smaller mutual funds with lower AUM figures are less diversified than larger mutual funds. The sub-optimal diversification can eventually explain a material portion of the future underperformance of these funds, if the return patterns are determined actually by fund managers' inability to replicate prior lucky returns, while fund investors are incapable of observing that despite their benchmarking attempts. According to our best knowledge, it seems that nobody has never either tested or controlled for the previously described alternative causality in an appropriate manner.

2.2 Mutual fund benchmarking in inefficient market

Siegel, Waring & Scanlan (2009) provide a concise encapsulation of the principles behind every investment decision. Three principles in particular are of great significance from the perspective of this thesis. The principles are related to accurate performance evaluation of fund managers, highlighting its complexity from the perspective of common mutual fund investors.

The first principle is that any investment should be evaluated in terms of alpha and beta contributions to the total risk and return that each investment opportunity offers. Alpha is by definition a return that is attributable to skill, whereas beta return arises from common risk exposures that earn risk premiums. It is worth noting, that although alpha derived from econometric models using historical data is often interpreted as managerial skill, the figure is unable to directly separate actual skill from luck perfectly.

The second principle is that investors should consider fees and other costs in all investment opportunities, in addition to the return breakdown and risk exposures. In mutual funds' context, three important cost components arise. These are management fees, other operating expenses, and trading costs of the mutual funds.

The third principle is that alpha is a zero-sum game and tends to be a very expensive form of investing. When the market is not efficient, the skilled managers earn alpha at the expense of the less skilled. It is worth noting, that alpha is valuable for fund investors only if the managers are exceptionally skilled. Since alpha is a zero-sum game between fund managers, the average fund manager earns next to zero alpha gross of fees and other expenses. Given costs that the investors of the funds must bear, only truly extraordinarily skilled fund managers can beat the market consistently to the extent that fund investors are compensated as well.

Investors need to be systematically successful in identifying exceptionally skilled managers if they wish to benefit from that superior skill consistently. However, as skill in historical performance is difficult to distinguish from the luck that also translates to prior alpha estimates, a material source of misunderstanding, fund allocation bias, and overpaying for fund investments arise from the information asymmetry between fund managers and investors. The problem only becomes increasingly more amplified, if common factor exposures are also captured by the alpha coefficient estimates. The previously described particular inability to distinguish the true skill potentially makes fund investors subject to lower expected returns.

2.3 Persistence of benchmarking malfunctions

As described above, investors perform accurate benchmarking and performance evaluation only if they are capable of decomposing total gross returns into abnormal returns (alpha), luck, and beta exposures that earn risk premiums over time, while also taking into account fees and costs-related to the investments. Furthermore, not only should investors differentiate beta from other return components, but they also need to distinguish various sources of beta risk premiums, e.g., market, value, size, and cross-sectional momentum-related returns. Existing research based on mutual fund flows suggests, that this is not the case. In fact, a wide range of empirical evidence entails that the situation is quite the opposite. Mutual fund investors mistake signals of alpha for noise and *vice versa* continuously, which arise from financially unsophisticated benchmarking acumen.

Mutual fund flows are in the centre of attention when studies strive to infer investor behaviour. The logic behind it is an assumption that "*investors' mouths lie where their money is*", i.e., the costly undertaken signal represents accurately their true information interpretation processes. Thus, many of the prior work related to mutual funds focus on fund flow dynamics in the industry. A consistent theme across the previous work is that mutual fund flows exhibit a systematic and versatile range of imperfections, which have a material impact on the returns that investors are expected to gain from investing in the mutual funds according to their current principles.

The most fundamental finding that casts doubt over investor sophistication in benchmarking is the documented relationship between common FRRs and mutual fund flows. A wide range of academic research finds that common risk factors other than the market drive a substantial portion of mutual fund flows (Berk & van Binsberg, 2016; Barber, Huang & Odean, 2016; Chakraborty, Kumar, Muhlhofer & Sastry, 2018). According to these studies, investors appear to attend most to the market factor. All other factor-exposures are misperceived as alpha by the investors. Given how established roles size, value, and cross-sectional momentum factors have in explaining the variation in equity returns, it is interesting how consistently they are being ignored as a source of beta risk.

Apart from the above, further empirical evidence emerge regarding investor irrationality related to mutual fund flows. Friesen & Sapp (2007) find that from 1991 to 2004, poor timing of equity fund investors' decisions alone reduces the average returns of the investors by 1.56% *per annum*. Friesen & Sapp (2007) elaborate that the inability to time investments is most prevalent in both actively managed funds as well as index funds, indicating of mutual fund investors' return-chasing tendency.

Not only are mutual fund investors incapable of differentiating managerial skill from factor exposures or time their investments, but empirical evidence even implies that mutual fund investors can be plain fooled by cosmetics. Jain & Wu (2000) observe that advertised funds attract significantly more flows *ex-post* advertisement period, although their performance is not superior during the same time window. Sirri & Tufano (2002) find similar results by showing how mutual funds' ability to attract fund flows appears to be the most salient for funds that exert higher marketing effort, as measured by higher marketing expenses. Sirri & Tufano (2002) further show that fund inflows are also directly related to the size of the fund complex as well as the currently received media attention. Solomon, Soltes & Sosyura (2014) further find evidence on advertising and media coverage of mutual funds impacting investors' capital allocation across mutual funds. Cooper, Gulen & Rau (2005) as for show that mutual funds can attract an average cumulative abnormal inflow of 28% just one year after repositioning their fund to reflect some current hot investment style, e.g., value or growth.

Del Guercio & Tkac (2008) provide final hard-hitting evidence to strengthen the case against the rationality of mutual fund investors. According to their findings, Morningstar ratings have a substantial independent influence on investment allocation decisions among retail mutual fund investors. Ben-David, Li, Rossi & Song (2018) find similar patterns to exist almost a decade later, providing evidence of the consistency of the phenomenon over time.

Consistently, mutual fund flows are characterized by unsophisticated and even illogical investor behaviour. According to the previous literature in this field, investors can be influenced even by cosmetic effects. These findings challenge the ability of mutual fund investors to analyse factor-exposures of the mutual funds, and systematically distinguish them from managers' skill and luck. Therefore, these findings serve as a solid foundation for the BMH.

3 Proposed theory and hypothesis development

The above overview of prior literature gives us a fair reason to believe that mutual fund investors perform biased performance benchmarking activities. If mutual fund investors even fail to differentiate prior common factor exposures from more accurate alpha estimates systematically, we have no reason to believe that the investors are successful in even more difficult activity of decomposing the alpha estimates into luck and actual skill components. The previously described limitation in performance benchmarking can be a direct cause of the patterns documented by Song (2020). We refer to the above intuition as *The Benchmarking Misevaluation Hypothesis.*

As described in the introduction of this thesis, we believe that a significant portion of the earlier findings documented as decreasing economies of scale is a consequence of managers having difficulties to replicate their earlier lucky performance going forward. We hypothesise, that mutual funds with the highest FRRs over a certain estimation period are mutual funds with high common risk exposures that have realised in a favourable manner. Observing the performance of these mutual funds going forward, we believe that these funds underperform to their benchmark adjusted targets systematically, since the funds are sub-optimally diversified while the fund managers simply run out of luck.

To test the above intuition, we formulate five hypotheses for this thesis. Each hypothesis addresses some aspect of the conditions that we expect to prevail under the BMH framework. The hypotheses with brief elaborations are reported below.

Hypothesis 1: Inverse relationship between prior FRRs and future abnormal returns prevails even if fund share classes are not rewarded with excess fund flows by investors

Given that we hypothesise that the relationship between prior FRRs and future benchmark-adjusted fund performance is attributable to fund managers' inability to replicate lucky returns rather than funds growing beyond their managers' active management skills as Song (2020) proposes, we expect that the inverse relationship between prior high FRRs and future underperformance prevail among all funds despite share classes' fund flow patterns. We test this hypothesis by extracting mutual fund share classes that do not exhibit irrational fund flows both when measuring fund flows in relative to the size of the share classes and in absolute dollar value. If we find that share classes with the highest prior FRRs over a certain estimation period underperform in the future despite not attracting excess fund flows, we conclude that the evidence aligns with the BMH.

Hypothesis 2: Inverse relationship between prior FRRs and future abnormal returns is more prevalent among smaller and presumably less diversified mutual fund share classes

Since we expect that the documented inverse relationship is to some extent attributable to lucky risk realisations in the past that are not replicated in the future, we expect that the phenomenon is stronger among mutual fund share classes with lower earmarked TNA figures. As Pollet & Wilson (2008) observe implicitly, smaller mutual funds tend to be less diversified compared to the larger mutual funds. Sub-optimally diversified portfolios as for are often riskier, yielding possibly higher returns in case the risks realise in a favourable manner.

Hypothesis 3: Share classes with the highest prior FRRs do not display future return patterns that are equally explainable with factor exposures compared to prior returns

Among the main components of the BMH is that high FRRs are often attributable to luck rather than skill in factor timing performed by the managers. Therefore, we hypothesise that mutual fund share classes with the highest FRRs over a certain estimation period do not exhibit subsequent return patterns that are equally explainable with common risk factor exposures, such as size, value, and cross-sectional momentum. We test this hypothesis by regressing the return patterns of the prior estimation periods and subsequent evaluation periods, and compare the obtained coefficients with one another. We conclude that we find evidence in favour of the BMH, if we observe that factor coefficients in the regression analyses are smaller and less significant during the future evaluation periods compared to the preceding FRR-estimation periods.

Hypothesis 4: Mutual funds with less sophisticated investor bases exhibit amplified BMH patterns, as these investors are more prone for errors in decomposing funds' performance

Given that the problem arises from benchmarking malfunctions, we hypothesise that the effect is amplified among mutual funds whose investors are less sophisticated in performance benchmarking. We believe, that in addition to the inability to distinguish luck from actual skill, the less-sophisticated investors are prone to several other benchmarking-related biases, including realising the agency costs of mutual fund managers' excess risk taking, despite that mutual funds are obligated to report their holdings systematically. We hypothesise, that more financially sophisticated investors are capable of practicing more diligent governance over the fund managers' actions compared to less sophisticated by punishing the fund managers with withdrawals if the investors observe fund managers misbehaviour. Similarly, investors who are less competent in fund performance benchmarking are less likely to observe fund managers' excess risk taking and other agency costs, providing these fund managers a broader set of opportunities to misbehave as principal of their clients. Therefore, we hypothesise that those mutual funds with less sophisticated investors are more prone to display BMH-associated return patterns, and *vice versa*.

> *Hypothesis 5: Inverse relationship between prior FRRs and future abnormal returns diminishes when the differentiation of good managers from bad is less relevant*

If the relationship between prior FRRs and future fund performance is a direct consequence of investors' inability to distinguish repeatable returns from those that are not, then the observed magnitude of the phenomenon should be dependent on how imperative the successful benchmarking is at the time. We hypothesise that during the past decade characterised by lower idiosyncratic volatility, decreased cost of information acquisition, and substantial appreciation of equities due to the Quantitative Easing (QE) performed by the Federal Reserve System[6], the performance difference between funds with higher and lower prior FRRs tapers. Hence, if we find that the future performance difference between funds with higher and lower prior FRRs is smaller during the past decade, we conclude that we find evidence in favour of the BMH.

---

[6] See *Appendices* Figure A. I.

4 Data and methodology

4.1 Mutual fund datasets

Mutual fund data comes from The Center of Research in Security Prices (CRSP) Survivor-Bias-Free U.S. Mutual Fund Database. The sample set comprise monthly return data and quarterly fund summaries of the U.S. mutual funds between 1992–2019. We choose 1992 as the starting point for the sample, since prior to that mutual funds were not obligated to report their returns and NAV figures on a monthly basis. Furthermore, having 1992 as a starting point compromises little if any representability of the available data, since mutual fund data is unreliable before 1984 (Fama & French, 2010; Song, 2020), while the current sample continues to cover data for nearly thirty years.

We process the data with several filters to achieve the final sample. First, we filter out all other mutual funds except actively managed equity mutual funds, i.e., funds that have ED-starting four-letter code describing the investment style of the mutual funds in the database. We exclude other funds since we ought to work with data of mutual funds where managerial skill should matter in the performance patterns of these funds. We continue to perform similar filtering as Song (2020), i.e., we exclude all index funds, ETFs, balanced funds, and sector funds. Finally, we remove few additional funds which despite being registered as actively managed equity funds in the database, the primary strategy of these funds is related to other asset classes, including but not limited to currencies, commodities, and fixed income products.

We continue by filtering out all mutual funds that have never managed assets more than five million (the U.S. dollars), or that do not have available monthly return data for at least 48 months straight. Similarly, we filter out all mutual funds that have no quarterly summary information for at least four years straight. We justify the previously introduced filters with our chosen methodology, which requires data for at least four years straight. The methodology is common in asset pricing studies, starting from Fama & Macbeth's (1973) study on the CAPM model. Furthermore, we use similar filters as Song (2020) applied so that we maintain the comparability and complementariness of the findings.

The only difference between our sample formulation process compared to that of Song's (2020) is that we do not aggregate mutual fund share classes into single mutual funds. In other words, we treat each share class of the mutual fund as a single fund. Despite introduction of potential biases in instances where single mutual funds are broken down into separate share classes and all share classes inherit the return history of the original mutual fund, we justify this decision to extract the investor behaviour regarding their tendency to respond to FRRs with excess fund flows, as well as to study the subsequent performance *ex-post* the estimation period. Thus, we use the term *mutual funds* for both the complete fund entities and their respective share classes interchangeably from this point onwards.

4.2 Risk factor data

Factor data used in the benchmarking model is from Kenneth French's website at Dartmouth. The dataset comprises monthly returns of three Fama–French Portfolios, cross-sectional momentum (latest quarter), and a risk-free return. The three Fama–French Portfolios are the market-weighted portfolio, the size portfolio (SMB), and the value portfolio (HML). Unlike Song (2020) does, we do not orthogonalise the factors, but rather use them as they are available in the website.

4.3 Estimating abnormal returns

Measuring the average monthly benchmark-adjusted returns of the mutual funds, we use a common four-factor model (FFC) that Carhart (1997) used in his research on persistence in mutual fund performance, and which has been since used widely in relevant literature regarding the performance of the mutual funds. Thus, the benchmarking model takes the form of

$$(1)\quad r_{i,\tau} - r_{f,\tau} = \alpha_{i,t}^{FFC} + \beta_{i,t}(MKT_\tau - r_{f,\tau}) + \gamma_{i,t}SMB_\tau + \theta_{i,t}\,HML_\tau + \delta_{i,t}UMD_\tau + \varepsilon_{i,\tau},$$
$$\tau \in \left\{t - \frac{m}{12}, \ldots, t - \frac{1}{12}\right\}$$

where $r_{i,\tau}$ is the mutual fund i's monthly return net of fees and other operating expenses in month τ, and $r_{f,\tau}$ stands for the risk-free rate of return at month τ. MKT_τ is the return of the value-weighted market portfolio, while SMB_τ, HML_τ, and UMD_τ are the returns of the size, value, and cross-sectional momentum portfolios, respectively. $\beta_{i,t}, \gamma_{i,t}, \theta_{i,t}$, and $\delta_{i,t}$ are the fund exposures to each factor in the benchmarking model, respectively. $\varepsilon_{i,\tau}$ equals the noise term that is assumed to satisfy the ordinary least square (OLS) assumptions.

4.4 Estimating factor-related returns

Using the OLS estimates of $\gamma_{i,t}$, $\theta_{i,t}$ and $\delta_{i,t}$, derived from the above equation (1), we estimate the factor-related average monthly returns over the 48 months long estimation periods as follows:

$$(2)\ \Delta_{i,t} = \frac{1}{m}\sum_{\tau=t-\frac{m}{12}}^{t-\frac{1}{12}}\left(\gamma_{i,t}SMB_{\tau} + \theta_{i,t}HML_{\tau} + \ \delta_{i,t}UMD_{\tau}\right),$$

where m equals the length of the regression period, and τ stands for each month involved in the regression period t. We leave the market factor unconsidered, since market-related returns have been shown to not have statistically significant explanatory power for mutual fund flows (Berk & van Binsbergen, 2016; Barber, Huang & Odean, 2016).

As mutual fund investors tend to interpret all other factors than the market as alpha, we also estimate the CAPM alpha for each fund i over the same prior m months with a traditional one-factor market model satisfying the OLS assumptions:

$$(3)\ r_{i,\tau} - r_{f,\tau} = \alpha_{i,\tau}^{CAPM} + \beta_{i,t}\left(MKT_{\tau} - r_{f,\tau}\right) + \epsilon_{i,\tau},$$

where $\epsilon_{i,\tau}$ is the monthly noise term that is assumed to satisfy the OLS assumptions.

4.5 Estimating fund flows

The net accumulation of funds is ultimately a function of three drivers. These are the fund inflows, fund outflows, and generated return (loss) on total net assets under management. Previous research has used a similar methodology[7] in estimating monthly average fund flows. Following the previous research and above introduced intuitive logic, the estimated fund flows at time t are

$$(4)\ Flow_{i,t} = TNA_{i,t} - \prod_{m=0}^{2}\left(1 + r_{i,t-\frac{m}{12}}\right)TNA_{i,t-\frac{1}{4}}$$

[7] E.g., see Song (2020).

where $TNA_{i,t}$ is fund i's total TNA at time t.

4.6 Description of the final sample

We obtain the final dataset by winsorising the estimated values of FFC alphas, fund flows, and FRRs. We perform the winsorising for the most extreme 1% of the observations in both ends of the distribution to control for outlier estimates. We report the key descriptive statistics of the sample in Table I below.

Table I
Descriptive statistics of the final sample: 1992–2019

This table displays the descriptive statistics of the final sample. The statistics and key ratios address the common characteristics that impact the performance of the mutual funds, which are the primary clientele, average TNA, average quarterly net flows, average monthly returns, volatility of returns, skill (α^{FFC}), average annual portfolio turn (turn ratio), and average annual expense ratios of the funds.

Panel A: Clientele of the sample	1992–1999	2000–2009	2010–2019
Number of funds, total	2,379	9,258	13,624
of which retail funds	1,839	6,412	6,519
of which institutional funds	467	2,695	5,454
of which both institutional and retail funds	73	151	1,651
Panel B: Performance metrics of the sample	**1992–1999**	**2000–2009**	**2010–2019**
Avg. TNA/share class (USD million)	1,081	621	711
Median TNA/share class (USD million)	190	77	80
Avg. quarterly flows/share class (USD million)	8.9	1.9	-2.9
Net returns, TNA-weighted average (%)	24.11	3.76	12.25
Annualised monthly volatility (%)	21.81	56.28	14.15
Sharpe ratio	1.11	0.07	0.87
Annualised monthly alpha (FFC) (%)	-1.93	-0.34	-1.50
Average annual turn ratio (%)	0.85	0.94	0.76
Median annual turn ratio (%)	0.66	0.67	0.47
Average annual expense ratio (%)	1.33	1.43	1.21
Median annual expense ratio (%)	1.23	1.33	1.15

5 Results

5.1 The relationship between FRRs and excess fund flows

In this section, we confirm to some extent that mutual fund investors confuse fund managers' skill for common beta exposures that are supposed to earn risk premiums over time. The findings align with the earlier results reported by Song (2020), as well as those of Berk & van Binsberg (2016), Barber, Huang & Odean (2016), Ben-David, Li, Rossi & Song (2018), and Chakraborty, Kumar, Muhlhofer & Sastry (2018). However, we also provide evidence of that the earlier findings related to mutual fund investors' irrational capital allocation between funds according to prior FRRs are not robust over the relative size of the mutual funds' share classes measured in TNA. In other words, it seems that the majority of the difference in excess fund flows attracted by FRRs vanish once we perform tests for differences in mean of average quarterly fund flows as a percentage of total net assets at the start of each estimation period.

To demonstrate the relationship between FRRs and fund flows, we construct fifteen fund portfolios of mutual fund share classes according to their past four-year abnormal returns and FRRs. We analyse the relationship between fund flows and FRRs by adopting a rolling window approach. For each calendar year of data, all mutual funds are sorted into five quintile portfolios based on their FFC alpha over the prior 48 months. The FFC alpha is a proxy for the skill of the managers. We then divide all funds into three portfolios according to the full sample distribution of average monthly FRRs over the same four-year time period. Thus, the top, middle, and bottom tercile portfolios represent portfolios of mutual funds whose estimated factor-related average past 48-month returns are in the top, middle, and bottom third of all mutual funds during their respective estimation periods.

Having divided share classes into fund portfolios, we compute the mean TNA at the start and at the end of each estimation period, continuing to compute the average quarterly net flows over the respective time periods. We then compute the annualised average time series CAPM alphas for each fund portfolio. Finally, we perform a two-way t-test of difference in means of average quarterly net fund flows between the top and bottom tercile fund portfolios. We report the relationship between FRRs and fund flows, as well as other results derived from the above computations in Table II.

Four out of five top tercile portfolios grow on average during the four-year estimation periods. The earmarked TNA figures range between USD million 287–605 at the start of the

Table II
The response of mutual fund flows to factor-related returns: 1992–2019

For each calendar year of data, all mutual funds are sorted into five quintile portfolios based on their four-factor alpha in prior 48 months. Each alpha-portfolio is then divided into three tercile portfolios based on their order in sample distribution of factor-related returns during the full sample period. Thus, the top, middle, and bottom tercile portfolios represent a portfolio of mutual funds whose estimated factor-related average past 48-month returns are in the top, middle, and bottom third of all mutual funds. A_s and A_e (millions in U.S. dollars) are the time series averages of cross-sectional arithmetic means of total net assets under management for a given fund portfolio at the start and the end of each 48-month estimation period. Flow (in millions) is the time series averages of the cross-sectional mean of average quarterly flows over each same 48-month estimation periods. α^{CAPM} and Δ (in %) are the annualised time series averages of the cross-sectional mean of the CAPM alphas and factor-related returns, respectively. t-stat is the test statistic of a test of difference in average quarterly flows between the top and bottom tercile portfolios within a given quintile, and t-stat^{RELA} performs similar test in difference of means, but accounting for the size of the funds. *, **, and *** denote 10%, 5%, and 1% statistical significance, respectively.

	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	t -stat	t-stat^{RELA}
	Top Tercile Δ					Middle Tercile Δ					Bottom Tercile Δ					Difference	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1 (L)	287	246	-4.6	3.6%	-3.2%	346	254	-6.1	0.7%	-5.1%	374	220	-6.3	-1.7%	-8.8%	4.50***	-1.73*
2	343	403	-2.8	3.0%	0.1%	395	403	-3.5	0.6%	-2.0%	514	414	-6.5	-1.3%	-4.0%	7.77***	0.87
3	432	628	-0.8	2.5%	1.0%	595	741	0.1	0.6%	-0.6%	706	743	-2.0	-1.3%	-2.5%	2.08**	-0.96
4	605	1,034	2.0	2.7%	2.4%	820	1,239	2.7	0.7%	0.8%	792	1,008	0.7	-1.4%	-1.3%	2.25**	-0.06
5 (H)	448	890	6.6	3.5%	7.4%	633	1,151	8.4	0.8%	3.9%	532	816	5.1	-1.9%	2.2%	2.64***	-0.96

estimation periods, reaching USD million 246–1,034 by the end of them. The only quintile whose funds do not grow on average is the bottom quintile, which comprise mutual funds with the worst fund managers estimated by the FFC alpha.

The portfolio that attracts the highest average quarterly fund flows is also the one with the highest abnormal returns measured with one-factor market model (CAPM alpha), while those funds with the lowest CAPM alpha have the least net fund flows. The average quarterly net flows are between USD million -4.6 and 6.6 across all quintiles, with corresponding CAPM alphas between -3.2% and 7.4%, respectively. The increasing trend in both CAPM alphas and fund flows when advancing from first quintile towards the fifth clearly demonstrates that mutual fund investors punish those mutual fund share classes with the lowest CAPM alphas with capital withdrawals. However, not all growth is attributable to higher fund flows, as we can see from the column 3. Another driver behind the growth of the earmarked TNA is the generated return for the held assets over the observation periods, since only two fund portfolios out of five exhibit positive net fund flow patterns.

Middle tercile portfolios demonstrate similar patterns in the respective measures as did top tercile peers, with minor differences in details. First, four out of five quintiles' share classes grow on average over the four-year observation periods. In these fund portfolios, mutual funds have average TNAs ranging between USD million 346–820 at the start of the estimation periods, reaching USD million 254–1,239 by the end of the periods, depending on the market-adjusted performance. It seems that all quintiles in the middle tercile exceed those in top tercile measured in TNA both at the start and at the end of the estimation periods, which is a different pattern compared to what Song (2020) reports with aggregated mutual fund estimates. All fund portfolios with the highest average time series CAPM alphas continue to attract the most fund flows, while lower average quarterly flows appear to follow directly the change in magnitude of the estimated CAPM alphas.

Bottom tercile funds are no exception. With an average earmarked TNA of USD million 374–792 at the start of the benchmarking periods and USD million 220–1,008 at the end of them, quarterly fund flows seem to continue following CAPM alphas. Interestingly, the size of the fund portfolios' average TNA at the end of the estimation periods are almost identical with the top tercile funds. When it comes to CAPM alphas, the figures range between -8.8% and 2.2%. Of these figures, four out of five fund portfolios (quintiles 1–4) have negative one-factor market-adjusted returns. We observe similar patterns with the average quarterly net fund flows, which fall in the range of USD millions -6.3 and 5.1. Again, four fund portfolios out of five demonstrate negative net flows. The flow estimates of the bottom tercile portfolios appear to be materially smaller compared to the corresponding figures in the top tercile fund portfolios.

To test if top tercile funds attract more fund flows compared to bottom tercile funds with similar level of exhibited skill, we perform a two-way test of difference in means (t-test) of average quarterly net fund flows. The test statistics of each quintile range between 2.07–7.77, demonstrating statistically significant (1–5% significance level) and consistent evidence of that fund portfolios with higher FRRs have systematically higher quarterly average net fund flows measured in absolute dollar value. The direct implication of the findings is that mutual funds with higher FRRs fool investors to perceive them more skilled than they actually are, instigating a bias in fund allocation decisions made by fund investors.

However, when we account for the average size of the funds at the start of the four-year estimation periods, the difference in means of net fund flows relative to the size of the funds is not significantly different between top and bottom tercile portfolios. Performing a similar test of difference in means, the test statistics range between -1.73–0.87, being all insignificant at 5% level. It seems that the majority of the discrepancy in excess fund flows between top and bottom tercile fund flows can be explained by the TNA of the funds' share classes, either due to the ticket size of the fund investors or due to the breadth of the investor base behaving in a similar manner to one another.

The above findings reveal three important details that have significant implications for the BMH. First, unlike in Song's (2020) paper with aggregated mutual funds, share classes with the highest prior FRRs do not grow to become the largest share classes on average measured in earmarked TNA. Instead, top tercile fund portfolios seem to become at best as large as bottom tercile funds by the end of the four-year estimation periods, falling short to the middle tercile fund portfolios considerably. This is an important empirical insight, since the observation itself already contradicts some of the decreasing economies of scale patterns that Song (2020) reports in his paper. Second, it seems that the share classes with the highest prior FRRs are consistently smaller mutual funds on average. The relevance of this empirical observation stems from that the pattern aligns with the BMH from the perspective of sub-optimal diversification that dictates which share classes transpire among the top third of all mutual fund share classes measured in FRRs. Third, when controlling for the size of the mutual funds' share classes at the start of the four-year observation periods, no excess fund flows attributable to FRRs appear to occur. The relevance of this observation as for arises from the fact that unlike Song (2020) suggests, the inverse relationship between prior FRRs and future benchmark-adjusted performance is not necessarily connected with the decreasing economies of scale hypothesis, but rather with the fund managers' inability to replicate lucky returns in the future that have been attributable to previously sub-optimal diversification and hence risk-taking.

The above insights set the foundation for analysing the BMH, as we already control for the decreasing economies of scale patterns. Therefore, if we obtain similar results from analysing the prior FRRs and the future fund performances as Song (2020) does, we find evidence in favour of the BMH. During the following sub-sections, we test the relationship between prior FRRs and future benchmark-adjusted performance and the Hypotheses 1–5 of this thesis, keeping in mind the above preliminary findings and conclusions.

5.2 Prior FRRs and future abnormal returns – lack of diversification and luck?

In this sub-section, we test the Hypotheses 1, 2, and 3 of this thesis by studying the performance of the fund portfolios both during and after the high FRR periods. We examine the relationship between mutual fund share classes' prior FRRs and future performance measured both in absolute net excess returns and abnormal returns compared to a proper benchmark. We then compare the factor composures of average monthly returns during the FRR estimation periods and their subsequent future performance evaluation periods. We find both favourable patterns yet inconclusive evidence of that prior FRRs are linked to future performance differences. Since the evidence is not entirely consistent, we conclude that the findings align significantly with the BMH but does not disposition earlier conclusions regarding the decreasing economies of scale in mutual funds.

We construct fifteen fund portfolios that we rebalance on an annual basis. Adopting a rolling window methodology, for each calendar year mutual funds are sorted into five quintiles based on their earmarked TNA at the start of the 12-month evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of prior four-year FRRs. Thus, the top, middle and bottom tercile portfolios within each quantile comprise mutual funds whose estimated factor-related average returns are among the top, middle, and bottom third of all mutual funds in that measure over the past four years.

We continue to compute the average TNA weighted monthly net excess returns of the fund portfolios for the next twelve months, labelled as the performance evaluation period. Thus, we compute a total of fifteen time series of monthly TNA-weighted net excess return patterns between 1996 and 2019. We then continue to regress these monthly average returns in excess of risk-free rate against the FFC model to obtain future benchmark-adjusted performance for the fund portfolios, as Song (2020) does in his work. We report the obtained results in Table III, in which the Panel A focuses on the net excess returns while Panel B displays benchmark-adjusted returns and standard errors of the FFC alpha estimates.

The reported results unveil some interesting patterns. First, FRRs and future fund performance measured in investors' net excess returns appear to be inversely correlated with one another, as top tercile portfolios generate on average 56 bps lower returns compared to their bottom tercile peers. The difference can be up to 70 bps when comparing the differences on fund portfolio level. Second, as reported in the Panel B, we find evidence that funds with higher prior FRRs demonstrate lower and statistically significant negative risk-adjusted returns in the future. On an aggregate level, top tercile fund portfolios have benchmark-adjusted returns of approximately 158 bps (1% significance level) during the 12-month evaluation period subsequent to the respective FRR estimation period. The figure falls short to both middle tercile portfolios at 83bps (1% significance level) and bottom tercile correspondent that do not exhibit significantly different estimates from zero. Furthermore, it seems that the phenomenon is more prevalent among smaller quintiles, suggesting that the sub-optimal diversification and managers' inability to replicate lucky returns could be the driving force in the reported patterns.

These findings align with the BMH to an extent. Given, that the phenomenon is more prevalent among smaller fund portfolios measured in average earmarked TNA, the evidence suggests that rather than exceeding the skill level of fund managers in managing actively larger funds, the potentially sub-optimal diversification of the holdings have substantial explanatory power over the future return patterns. Combining these findings with insignificant difference in size-adjusted average quarterly net fund flows earlier, we conclude to find evidence that is consistent with the Hypotheses 1 and 2 of this thesis. We obtain similar results when we control for skill instead of the size of the fund portfolios. (*Appendices*, Table A.I).

A key element of the BMH is that fund managers who earn FRRs among the top third of all funds do so primarily due to luck, and that the fund managers of these share classes are incapable of replicating the prior performance going forward. Empirically, we showed earlier that this phenomenon appears to be particularly present among smaller mutual fund share classes at the start of the 12-month evaluation periods, suggesting that the root cause could be sub-optimal diversification and lucky realisations of factor exposures in the past. To confirm the above interim conclusions, we compare the factor return exposures of the fund portfolios between the 48-month estimation periods and their subsequent 12-month future performance

Table III
Future annualised net returns, controlled for fund TNA: 1996–2019

For each calendar year, mutual funds are sorted in two dimensions. First, mutual funds are sorted into five quintile portfolios based on their total net assets under management at the start of the 12 months long evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the prior 48 months-long period. Thus, the top, middle and bottom portfolios within each quantile comprise mutual funds whose estimated factor-related average returns during the past 48 months are among the top, middle, and bottom third of all mutual funds in that measure. We then compute the average TNA weighted monthly net returns of each fund portfolio in excess of risk-free rates over the 12-month evaluation periods and annualise them. We continue by regressing these monthly average returns in excess to risk-free rate against the FFC model to obtain future benchmark-adjusted returns for each fund portfolio. Both Ret and α_{FFC} are annualised figures in percentages, while *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Figures in parenthesis are standard errors of OLS estimates.

	Panel A: Ret (%)				Panel B: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	6.55	6.30	6.53	-0.02	-1.91**	-1.14**	-1.11	0.80
					(0.90)	(0.56)	(0.88)	(1.41)
2	6.67	5.99	6.29	-0.38	-1.81**	-1.51***	-1.38	0.43
					(0.83)	(0.53)	(0.92)	(1.41)
3	7.08	6.77	7.14	0.06	-1.31	-0.62	-0.59	0.72
					(0.91)	(0.51)	(0.90)	(1.46)
4	6.87	6.91	7.40	0.53	-1.65*	-0.61	-0.39	1.26
					(0.88)	(0.56)	(0.93)	(1.49)
5(H)	6.92	7.01	7.62	0.70	-1.22	-0.27	0.11	1.33
					(0.74)	(0.40)	(0.95)	(1.40)
All	6.93	6.99	7.50	0.56	-1.58***	-0.83***	-0.67	0.91
					(0.39)	(0.24)	(0.41)	(0.65)

evaluation periods. We perform the comparison on top tercile fund portfolios, since they display the most substantial underperformance in the future measured in FFC alpha.

Again, we adopt a rolling window approach. For each calendar year of data from 1996 onwards, we sort mutual funds into five quintile portfolios based on their earmarked TNA at the start of the 12-month evaluation period. We then divide quintile portfolios into three terciles according the full sample distribution of FRRs over the prior 48 months long estimation periods. Thus, mutual funds in the top tercile portfolios are share classes with the highest FRRs over the prior four years. For only the top tercile funds, we continue to compute time series of the average monthly TNA weighted net excess returns over the preceding estimation periods and regress the return patterns against the FFC factors to obtain and report the return compositions of the fund portfolios prior to the evaluation periods. To control for that we obtain four TNA-weighted return observations for each month 1995 onwards due to overlapping estimation periods, we apply a panel regression model with controls for fixed effects[8] of the 48-month estimation periods. We then compute the average monthly TNA weighted net excess returns of the same top tercile fund portfolios over the next 12-month evaluation periods, but instead of performing a panel regression we perform a common OLS time series regression against the FFC risk factors, since we do not have the same overlap problem as we had with the estimation periods' returns. We report the results in Table IV.

The findings of the analysis align substantially with the assumption of that previous lucky realisations of high FRRs could explain a significant portion of the mutual fund share classes underperformance *ex-post* the FRR estimation periods. Observing the Panel A, we find that all three common FFC factor coefficients are significant (1% significance level) during times when share classes transpire in top third of all share classes measured in FRRs. SMB coefficients are between 0.23–0.38, HML coefficients are between 0.07–0.15, and UMD coefficients are between 0.08–0.11.

However, observing the corresponding figures in Panel B, i.e., factor exposures of the fund portfolios following the FRR estimation periods, we discern considerable changes both in magnitude and in significance of the factor exposure coefficients. Apart from the UMD, all factor coefficients fall from estimation periods. While SMB coefficients fall 0.05–0.09 units, HML coefficients decrease up to 0.19 units. Additionally, the value factor loses material portion of its prior significance, as none of the top tercile fund portfolios have significant coefficients for HML during the future performance evaluation periods. The evidence suggests that mutual fund share classes that belong to the top tercile measured in FRRs over a certain estimation period achieve that apparently by having higher and better-timed exposures to value and size, which they are not capable of sustaining during the subsequent periods of time. Thus, the evidence suggests the factor timing is attributable to luck, which aligns with Hypothesis 3.

In addition to the above patterns that are favourable for the BMH, we find further evidence that aligns with the Hypothesis 2 of this thesis. Observing the extent of BMH-associated patterns that each fund portfolio in the top tercile of FRRs display, it appears that mutual fund share class portfolios with the lowest TNA figures at the start of the evaluation period display the most material factor exposure changes between the two time intervals. For instance, looking at the quintile that denotes for the smallest fund portfolios measured in TNA

[8] The results are consistent also with pooled panel regression and panel regression with controls for random effects.

Table IV
Coefficient estimates of the top tercile funds' return components, 1992–2019

For each calendar year, mutual funds are sorted into quintile portfolios based on their earmarked TNA at the start of the 12-month evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns over the prior 48-month estimation period. We continue to compute the average TNA weighted monthly net excess returns of the estimation periods and regress them against the FFC model to obtain risk factor exposures of each fund portfolio at the time of the fund portfolio formation interval (Estimation period). To control for several estimation per month due to the overlapping estimation periods, we perform a panel regression controlling for fixed effects of the 48-month estimation intervals. We report the coefficients together with their OLS standard errors (in parentheses) in Panel A. We then compute the average TNA weighted monthly net excess returns of the fund portfolios during the subsequent 12-month evaluation periods and regress them against the FFC model to obtain risk factor exposures of each fund portfolio in the future. We report the coefficients together with their OLS standard errors (in parentheses) in Panel B. *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively.

	Panel A: Estimation period (prior 48 months)			Panel B: Evaluation period (future 12 months)		
	SMB	HML	UMD	SMB	HML	UMD
	(1)	(2)	(3)	(4)	(5)	(6)
1(L)	0.23***	0.15***	0.08***	0.14***	-0.04	0.06***
	(0.015)	(0.017)	(0.010)	(0.019)	(0.021)	(0.013)
2	0.31***	0.07***	0.11***	0.26***	-0.02	0.09***
	(0.016)	(0.018)	(0.011)	(0.023)	(0.025)	(0.015)
3	0.36***	0.07***	0.10***	0.29***	-0.03	0.10***
	(0.013)	(0.014)	(0.008)	(0.024)	(0.026)	(0.016)
4	0.37***	0.08***	0.10***	0.30***	0.03	0.11***
	(0.013)	(0.014)	(0.009)	(0.022)	(0.023)	(0.014)
5(H)	0.38***	0.09***	0.10***	0.33***	0.02	0.11***
	(0.014)	(0.015)	(0.009)	(0.024)	(0.025)	(0.016)

(1(L)) we find that SMB and HML coefficients drop 0.09 and 0.19 between the two time intervals, respectively. The changes are the largest among all top tercile's quintile portfolios. It seems that mutual fund share classes with the lowest earmarked TNA-figures are also less diversified, since the risk exposure coefficients of their respective return patterns are the most volatile between the two time periods, suggesting that smaller share classes are more prone to the BMH-associated return patterns. Combining this insight with the earlier findings in favour of the Hypothesis 3, we find significant evidence suggesting that BMH prevails in the actively managed equity mutual funds in the U.S.

Notwithstanding, despite the significant evidence in favour of the BMH that we present in this sub-section, the evidence remains inconclusive. As we report in Table III, the difference in benchmark-adjusted future returns between top and bottom tercile funds are consistently insignificant. Hence, the above findings lack power to disposition any earlier findings related to the decreasing economies of scale hypothesis. However, the findings emphasise the lack of attention that the earlier research on decreasing economies of scale in the mutual funds' context has had on investors' benchmarking malfunctions until just recently. Including investors' inability to recognise that fund managers with the highest "alphas" tend to be lucky managers who are incapable of replicating the return patterns in the future.

5.3 Investor base and its role between funds' prior FRRs and future performance

In this section, we test Hypothesis 4 of this thesis by studying FRRs, the fund flow patterns, and future benchmark-adjusted performance differences between mutual funds that serve solely either institutional clientele or exclusively retail investors. We report that mutual funds with less sophisticated investor bases both by definition and empirically exhibit amplified inverse relationship between prior FRRs and future benchmark-adjusted returns. The evidence aligns with the BMH, showing that mutual fund investors with less sophisticated benchmarking skills are more prone to experience the BMH-associated return patterns that earlier have been labelled as decreasing economies of scale.

Moreover, we find evidence of that mutual fund share classes with more sophisticated investor bases exhibit similar but more moderate BMH-related return patterns, despite that more sophisticated investors do not exhibit tendency of rewarding fund managers' performance attributable to other variables than skill with excess fund flows. Unlike with full sample analysis earlier, we find that, e.g., institutional investors do not exhibit irrational fund flow patterns even when measuring net fund flows in absolute dollar value. Thus, we find further evidence that aligns with the Hypothesis 1 of this thesis in addition to the Hypothesis 4.

However, the differences in future performance between top and bottom tercile funds remain insignificant. Top and bottom tercile estimates do not exhibit any significant difference in future benchmark-adjusted performance patterns, despite that top terciles underperform their respective benchmarks significantly while bottom terciles do not – regardless of whether the researched mutual fund share classes are retail or institutional funds. Therefore, the evidence remains inconclusive in grand scheme of things.

We start by comparing the excess fund flow patterns between top and bottom tercile portfolios for both retail fund and institutional fund share classes that are associated with high prior FRRs. Adopting the rolling window approach, we sort all retail and institutional mutual funds into five quintile portfolios based on their FFC alpha during the prior 48 months. We then divide all funds into three portfolios according to the full sample distribution of factor-related returns over the same time period. Thus, the top, middle, and bottom tercile portfolios represent mutual funds whose estimated factor-related average past 48-month returns are in the top, middle, and bottom third of all mutual funds in their respective groups (retail or institutional).

Having divided funds into respective fund portfolios, we compute the mean TNA at the start and at the end of each estimation period, followed by computations of the average quarterly flows. As earlier, we also compute the annualised average time series CAPM alphas of the portfolios. Finally, we perform a two-way t-test of difference in mean between the top and bottom tercile fund portfolios' net fund flows. We report the relationship between FRRs and fund flows, as well as other results derived from the above computations in Table V.

The findings suggest that the less sophisticated mutual fund investors by definition appear to be also empirically less sophisticated in benchmarking, when inferring their financial sophistication from fund flow patterns. While all retail fund portfolios display significant (1% significance level) excess fund flows patterns between top and bottom tercile portfolios, institutional fund portfolios exhibit no significant differences at all. In fact, presumably professional institutional investors are so competent in differentiating common factor exposures that all corresponding fund portfolios in all Terciles have almost identical flow patterns within respective skill quintiles. The fund flows are of similar order of magnitude both in absolute dollar value as well as normalised against the size of the funds. Thus, it seems that institutional investors are not just more sophisticated by definition, but they actually behave empirically in a more sophisticated manner in allocating their funds compared to the retail investors.

An interesting observations that we make from the Table V is that the TNA figures of the retail funds at the start of the estimation periods are on average materially larger than their corresponding fund portfolios of institutional funds. In fact, the difference in size is so substantial that some retail fund portfolios are almost twice as large as those institutional mutual funds in corresponding fund portfolios. This insight is important going forward when analysing the difference in future benchmark-adjusted performance, as some of the performance differences can be attributable to decreasing economies of scale due to exceeding the active management skills of the respective fund managers. However, we also note that both control groups, and particularly the retail fund portfolios, exhibit similar patterns of that smaller funds at the start of the fund flow estimation period appear to end up in the top tercile. This insight

Table V

Fund flows and factor-related returns, retail vs. institutional funds: 1992–2019

We start by dividing mutual fund share classes into retail and institutional fund groups. For each calendar year of data, mutual fund groups are sorted into five quintile portfolios based on their four-factor alpha during prior 48 months. Each alpha-portfolio is then divided into three tercile portfolios based on their order in sample distribution of factor-related returns over their respective 48-month estimation periods. Thus, the top, middle, and bottom tercile portfolios represent a portfolio of mutual funds whose estimated factor-related average past 48-month returns are in the top, middle, and bottom third of all retail mutual fund share classes in a given group. A_s and A_e (millions in U.S. dollars) are the time series averages of cross-sectional arithmetic means of total net assets under management for a given fund portfolio at the start and the end of each 48-month estimation period. Flow (in millions) are the time series averages of the cross-sectional mean of average quarterly flows over each same 48-month estimation periods. Δ and α^{CAPM} (in %) are the annualised time series averages of the cross-sectional mean of the factor-related returns and CAPM alphas, respectively. t-stat is the test statistic of a test of two-way difference in average quarterly flows between the top and bottom tercile portfolios within a given quintile. *, **, and *** denote 10%, 5%, and 1% statistical significance, respectively. Quintiles 1–5 represent retail fund share classes, while 1'– 5' stand for institutional fund share classes.

	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	t-stat
	Top Tercile Δ					Middle Tercile Δ					Bottom Tercile Δ					Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1 (L)	302	239	-5.3	4.0%	-3.4%	366	243	-6.8	0.8%	-5.6%	418	219	-7.4	-2.0%	-9.8%	4.55***
2	361	435	-2.8	3.3%	0.1%	455	410	-5.5	0.7%	-2.1%	584	400	-8.7	-1.5%	-4.4%	9.79***
3	510	724	-1.6	2.9%	1.2%	695	816	-1.8	0.7%	-0.6%	854	816	-4.5	-1.4%	-2.7%	4.25***
4	656	1,094	1.8	3.1%	2.9%	1,043	1,543	2.0	0.8%	1.0%	1,002	1,181	-0.6	-1.5%	-1.4%	3.14***
5 (H)	527	1,064	7.4	4.0%	8.3%	725	1,297	8.7	0.9%	4.3%	651	947	4.7	-2.1%	2.6%	3.67***
1'(L)	235	196	-3.7	3.3%	-2.3%	240	198	-3.7	0.6%	-4.2%	301	225	-4.4	-1.4%	-7.2%	1.17
2’	257	307	-1.9	2.7%	0.2%	331	399	-1.1	0.6%	-1.6%	358	374	-2.0	-1.2%	-3.5%	0.04
3’	292	469	1.5	2.2%	0.8%	381	526	2.4	0.5%	-0.5%	511	665	2.4	-1.2%	-2.2%	-1.06
4’	549	968	2.8	2.2%	2.1%	539	908	4.7	0.6%	0.8%	530	832	4.1	-1.2%	-1.1%	-1.28
5 '(H)	325	692	7.7	2.9%	6.6%	427	832	8.8	0.8%	3.6%	298	561	6.8	-1.6%	2.3%	1.03

as for aligns with the BMH, strengthening the hypothesis that links the future benchmark-adjusted performance with fund managers' inability to replicate historical lucky risk realisations during the subsequent performance evaluation periods.

To compare the differences in patterns between future benchmark-adjusted returns and prior FRRs, we perform similar analysis for both retail funds and institutional funds, as we performed for the full sample in Table III. Adopting the 48-month rolling window methodology, first we sort all mutual funds into five quintiles based on their TNA at the start of the 12-month performance evaluation period. We then divide all TNA quintile portfolios into three tercile portfolios based on the sample distribution of prior four-year FRRs. We continue to compute the average TNA weighted monthly net excess returns for the fund portfolios over the 12-month evaluation periods, deriving a total of thirty time series of monthly TNA-weighted return patterns between 1996 and 2019. We continue to regress these monthly average returns in excess of risk-free rate against the FFC model to obtain future benchmark-adjusted return estimates for the fund portfolios. We report the obtained results in Table VI, in which the Panel A and B focus on the performance of the retail funds, while Panels C and D report corresponding figures for the institutional funds.

From perspective of the legitimacy of the BMH, we find two interesting patterns. First, we find significant evidence (1–5% level) of that both retail and institutional funds associated with higher prior FRRs underperform their respective risk-adjusted benchmarks in the future. Top tercile fund portfolios of retail funds underperform their relevant benchmark by 181bps (1% significance level), while the corresponding figure for the institutional funds is only 91bps (5% significance level). Second, we find that the order of magnitude, consistency, and significance of the underperformance among retail funds seems to be higher compared to the institutional funds. Of fifteen retail fund portfolios, five of them display statistically significant and increasingly negative future benchmark-adjusted performance as prior FRRs increase. Meanwhile, among institutional fund portfolios only one fund portfolio exhibits statistically significant underperformance in the future.

The above findings are interesting for three reasons. First, it seems that those investors with provenly (Table V column 16) more severe troubles in differentiating luck, skill, and common factor exposures from each other encounter more severe penalty for their financial unsophistication in a form of lower future returns. This observation aligns with the BMH, and more precisely, the Hypothesis 4 of this thesis. Second, again, it seems that the inverse relationship between prior FRRs and future performance is more prevalent for funds that were

Table VI
Future annualised net returns of retail and institutional funds: 1996–2019

For each calendar year, mutual funds are sorted into quintile portfolios based on their earmarked TNA at the start of the 12-months evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the prior four years. We then compute the average TNA weighted monthly net returns in excess of risk-free rates for the 12-month evaluation periods and annualise them. We continue to regress these monthly average returns in excess to risk-free rate against the FFC model to obtain future benchmark-adjusted returns for each fund portfolio. Both Ret and α_{FFC} are annualised figures in percentages, while *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Parentheses report OLS standard errors. Panels A and B address retail funds, while panels C and D report institutional funds' corresponding figures.

	Panel A: Ret (%)				Panel B: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	5.91	6.09	6.22	0.31	-2.41**	-1.29**	-1.40	1.02
					(0.93)	(0.58)	(0.90)	(1.45)
2	6.42	5.74	6.10	-0.32	-2.00**	-1.66***	-1.45	0.56
					(0.85)	(0.54)	(0.96)	(1.44)
3	6.56	6.54	6.61	0.05	-1.81**	-0.84	-1.09	0.72
					(0.88)	(0.51)	(0.90)	(1.45)
4	6.92	6.68	7.30	0.38	-1.62*	-0.77	-0.42	1.20
					(0.89)	(0.57)	(0.98)	(1.52)
5(H)	6.91	7.11	7.70	0.79	-1.22	-0.20	0.21	1.43
					(0.74)	(0.40)	(0.97)	(1.41)
All retail	6.92	7.07	7.56	0.64	-1.81***	-0.95***	-0.83*	0.99
					(0.39)	(0.24)	(0.43)	(0.66)

	Panel C: Ret (%)				Panel D: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	7.82	7.08	7.34	-0.48	-0.82	-0.38	-0.36	0.46
					(0.90)	(0.64)	(0.92)	(1.37)
2	7.38	6.73	7.38	0.00	-1.14	-0.82	-0.59	0.55
					(0.85)	(0.54)	(0.88)	(1.43)
3	8.08	7.35	8.23	0.15	-0.35	-0.17	0.53	0.88
					(1.16)	(0.60)	(1.16)	(1.73)
4	6.67	8.55	7.51	0.85	-1.59**	0.49	-0.56	1.03
					(0.80)	(1.85)	(0.92)	(1.39)
5(H)	7.53	6.94	7.30	-0.23	-0.65	-0.54	-0.30	0.34
					(0.83)	(0.76)	(0.79)	(1.32)
All inst.	7.44	7.18	7.45	0.01	-0.91**	-0.28	-0.25	0.65
					(0.42)	(0.45)	(0.43)	(0.68)

the smallest at the start of the fund flow estimation periods, aligning with the Hypothesis 2 of this thesis. Third, despite low cross variation in fund portfolios' TNA between top and bottom tercile funds, as well as rational fund flow patterns across all quintiles, institutional funds with the highest prior FRRs also seem to underperform their relevant risk-adjusted benchmarks in the future. This observation further strengthens the Hypothesis 1 of this thesis. Thus, we conclude that the evidence is in favour of the BMH.

However, as it is the case with the results that we obtain from the full sample analysis, the evidence remains inconclusive. No matter how consistently and significantly negative future performance prevails among the retail funds' top tercile portfolios, we obtain no evidence of that future abnormal returns between top and bottom tercile funds differ significantly. Hence, the true nature of the difference remains questionable. To conclude the section, we find some evidence in favour of the BMH, but in grand scheme of things, the BMH requires further testing.

5.4 The BMH-associated patterns and favourable market conditions

Another reasoning that we make about the BMH is that the financial environment in which investors operate could impact the relationship between prior FRRs and the extent of future underperformance that the share classes with the highest prior FRRs are expected to have. For instance, if differentiation of skilled managers from lucky ones becomes either easier or less relevant, then we would expect the difference in future performance between top and bottom tercile funds to narrow down. By less relevant, we refer to times when mutual funds have relatively similar performance regardless of the skill of the fund managers. In this section, we test that reasoning by more precisely testing the Hypotheses 5 of this thesis.

We define the above described time period to be characterised by the following three attributes. First, we expect that the share of idiosyncratic volatility of the mutual funds' performance is lower compared to the full sample average. Second, the overall volatility of the return patterns are not on amplified levels compared to the full sample averages. Third, the U.S. equity markets are encountering a long and steady bull market, which preferably are further smoothened by regenerative monetary policy.

To detect a favourable time period for the above purpose, we start by performing an analysis of funds' average share of idiosyncratic volatility in total variation of the share classes' return patterns. Using a rolling-window methodology, we estimate the portion of idiosyncratic volatility in return patterns for each mutual fund share class over the past 48 months from

January 1996 onwards. We perform the estimations on a monthly basis by computing the cross-sectional average of monthly $1 - R^2$ figures of the mutual funds' time series regressions as described by the equation 1. We report the pattern of idiosyncratic volatility in Figure I. Observing the time series pattern, the sub sample that we identify as fitting from the perspective of idiosyncratic volatility is anything between 2008–2019. Despite that the share of idiosyncratic volatility is on the rise from 2008–2012 period onwards, it remains lower compared to the full sample average until 2016 and again 2018 onwards.

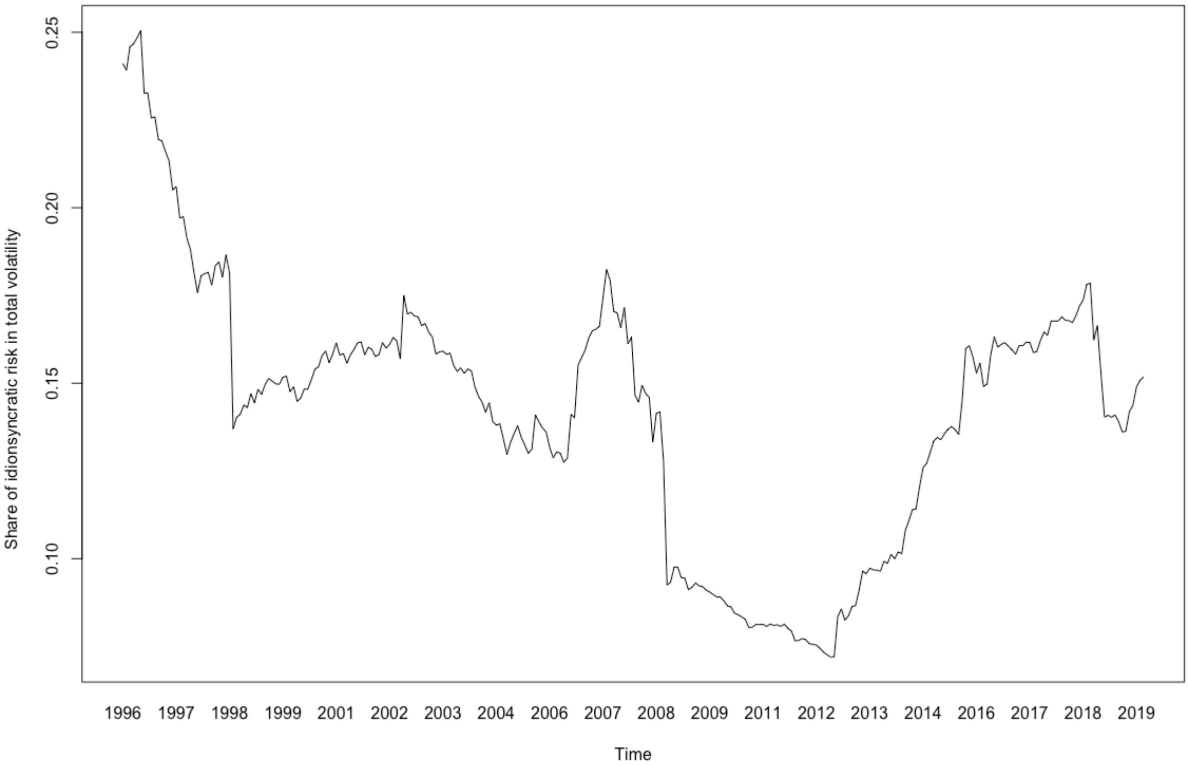
Combining the conclusions of the above analysis, the volatility measures reported in Table I of this thesis, and the monetary policy that prevailed in the U.S. at the time, we conclude that the latest full decade of 2010–2019 fulfils the three attributes that we discussed earlier in this sub-section. The time period serves the purpose for three reasons. First, as discussed above, the share of idiosyncratic volatility in total return variation is below its full sample average, while the time interval itself still remains long enough to perform a credible analysis. Second, the overall volatility of return patterns are far lower at approximately 14% *per annum* compared to over 20% and over 50% between 1992–1999 and 2000–2009, respectively. Third, the most recent decade is characterised by a historical period of exceptionally regenerative monetary policy practised by the FED ("the intervention") and other Central Banks around the world. Some believe that the intervention has resulted in increased correlation and depressed levels of volatility in many asset classes (Bollen, Joenväärä & Kauppila, 2020). To summarise, we expect that future performance difference between top and bottom tercile funds attributable to the BMH should narrow down during the past decade, since observing the truly skilled managers from lucky ones is arguably less relevant when all asset classes and particularly public equities have inflated due to the intervention of the FED[9] in a low volatility environment.

To test the Hypothesis 5 of this thesis, we perform two analyses in line with Tables III and V to study the impact of the market conditions both on an aggregate level as well as on a more granular level, accounting for the differences in benchmarking sophistication between the fund investors. The findings are reported in Tables VII and VIII. In grand scheme of things, the evidence aligns well with the framework of the BMH. First, looking at all the actively managed equity mutual funds of the sample over 2010–2019, almost all fund portfolios have negative and significant (1% significance level) future benchmark-adjusted returns *ex-post* FRR and fund flow estimation periods. Top tercile funds' future abnormal returns are -211bps (1% level),

[9] *Appendices* Figure A. I.

Figure I
Idiosyncratic volatility as a portion of the total volatility estimates: 1992–2019

This figure exhibits the average of estimated portion of idiosyncratic volatility in mutual funds' total volatility over the preceding four years. The portion of idiosyncratic volatility is estimated from the cross-sectional average of $1 - R^2$ figures from each mutual fund's time series regression over the prior 48 months. Adopting the rolling-window approach, we perform a similar estimation on a monthly basis to derive the time series of the idiosyncratic volatilities. To illustrate the methodology, e.g., the first observation of the graph in January 1996 denotes for the average share of idiosyncratic volatility of the mutual funds between January 1992 and December 1995.



while bottom tercile funds have almost as low a figure of -175bps (1% significance level). The evidence is consistent even at the fund portfolio level.

Interestingly, we observe similar patterns when we analyse the changes in the estimates on a more granular level in Table VIII. All retail and institutional fund portfolios have a negative and significant future benchmark-adjusted returns *ex-post* having belonged to the top tercile in FRRs over the preceding four-year estimation periods. The most interesting observation is within the institutional fund portfolios, where top and bottom tercile fund portfolios have almost identical estimates for the future FFC alphas, ranging between -165bps and -155 bps (1% significance level). The observation is particularly interesting, as it fits our hypothesis almost perfectly. Given that in full sample analysis, fund portfolios with the lowest prior FRRs did not exhibit significantly different performance estimates from zero, whereas during the 2010–2019 sub-sample period the bottom tercile exhibits significant future

Table VII
Future annualised net returns, controlled for fund TNA: 2010–2019

For each calendar year, mutual funds are sorted in two dimensions. First, mutual funds are sorted into five quintile portfolios based on their total net assets under management at the start of the 12 months long evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the prior 48 months-long period. Thus, the top, middle and bottom portfolios within each quantile comprise mutual funds whose estimated factor-related average returns are among the top, middle, and bottom third of all mutual funds in that measure during the past 48 months. We then compute the average TNA weighted monthly net returns of each fund portfolio in excess of risk-free rates over the 12-month evaluation periods and annualise them. We continue by regressing these monthly average returns in excess to risk-free rate against the FFC model to obtain future benchmark-adjusted returns for each fund portfolio. Both Ret and α_{FFC} are annualised figures in percentages, while *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Figures in parentheses are standard errors of OLS estimates.

	Panel A: Ret (%)				Panel B: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	10.21	9.36	9.68	-0.53	-2.46**	-2.03***	-2.16**	0.30
					(1.13)	(0.63)	(0.86)	(1.69)
2	10.50	9.36	9.73	-0.77	-2.30**	-2.19***	-2.19**	0.11
					(1.04)	(0.64)	(0.87)	(1.64)
3	10.76	9.81	10.07	-0.70	-2.12**	-1.87***	-1.75**	0.37
					(1.06)	(0.59)	(0.86)	(1.65)
4	11.05	10.14	10.28	-0.77	-2.04**	-1.64***	-1.67*	0.37
					(1.01)	(0.61)	(0.89)	(1.67)
5(H)	11.52	10.77	10.99	-0.53	-1.63*	-1.32**	-0.99	0.64
					(0.92)	(0.54)	(0.80)	(1.55)
All	11.45	10.70	10.89	-0.57	-2.11***	-1.81***	-1.75***	0.36
					(0.46)	(0.27)	(0.38)	(0.73)

benchmark-adjusted underperformance with similar order of magnitude to the top tercile funds, we find some evidence in favour of that the difference between the two terciles narrows down, and hence in favour of the Hypothesis 5 of this thesis.

However, it is essential to emphasise that the findings continue to remain inconclusive, although some additional evidence emerges in favour of the BMH. Since the difference between top and bottom tercile funds have been insignificant in all of the earlier analyses, we cannot draw any firm conclusions on that the gap between top and bottom tercile funds had in fact narrowed down significantly during the past decade. The above patterns that imply the convergence of the performance difference can be just semantics in the grand scheme of things. Having said that, we highlight that despite the results lack complete conclusiveness, again, the

Table VIII
Future annualised net returns of retail and institutional funds: 2010–2019

For each calendar year, mutual funds are sorted into quintile portfolios based on their earmarked TNA at the start of the 12-month evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns over the prior four-year period. We then compute the average TNA weighted monthly net returns in excess of risk-free rates over the 12-month evaluation periods and annualise them. We continue to regress these monthly average returns in excess of risk-free rate against the FFC model to obtain future benchmark-adjusted returns for each fund portfolio. Both Ret and α_{FFC} are annualised figures in percentages, while *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Parentheses report OLS standard errors. Panels A and B address retail funds, while panels C and D report institutional funds' corresponding figures.

	Panel A: Ret (%)				Panel B: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	9.19	8.89	9.17	-0.02	-3.29***	-2.39***	-2.44**	0.85
					(1.21)	(0.68)	(0.95)	(1.76)
2	9.97	9.11	9.47	-0.50	-2.65**	-2.34***	-2.19**	0.46
					(1.08)	(0.62)	(0.92)	(1.71)
3	10.53	9.39	9.76	-0.77	-2.35**	-2.17***	-1.95**	0.40
					(1.05)	(0.56)	(0.89)	(1.64)
4	10.90	9.77	9.81	-1.08	-2.34**	-2.00***	-1.92**	0.42
					(1.01)	(0.59)	(0.94)	(1.68)
5(H)	11.54	11.00	11.05	-0.49	-1.70*	-1.25**	-0.91	0.79
					(0.92)	(0.54)	(0.84)	(1.57)
All retail	11.46	10.90	10.92	-0.55	-2.47***	-2.03***	-1.88***	0.58
					(0.47)	(0.27)	(0.40)	(0.74)

	Panel C: Ret (%)				Panel D: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	11.49	9.74	10.49	-1.00	-1.53	-1.85***	-1.84**	-0.32
					(1.14)	(0.62)	(0.82)	(1.67)
2	11.19	9.83	10.36	-0.83	-1.62	-1.81***	-1.90**	-0.28
					(1.03)	(0.61)	(0.89)	(1.67)
3	11.03	10.17	10.40	-0.63	-1.84*	-1.78**	-1.47*	0.37
					(1.09)	(0.70)	(0.87)	(1.66)
4	11.19	10.53	10.96	-0.23	-1.67	-1.40**	-1.28	0.39
					(1.06)	(0.63)	(0.93)	(1.79)
5(H)	11.66	10.76	10.96	-0.70	-1.60	-1.52**	-1.25	0.35
					(1.00)	(0.59)	(0.85)	(1.66)
All instit.	11.61	10.74	10.94	-0.67	-1.65***	-1.67***	-1.55***	0.10
					(0.47)	(0.28)	(0.39)	(0.75)

evidence is significant and favourable to some extent. Furthermore, the evidence emphasises that the benchmarking malfunctions of fund investors should be controlled for in the future line of research regarding the performance of the mutual funds.

6 Robustness of the results

6.1 *Robustness of the Benchmarking Misevaluation Hypothesis*

Despite that the evidence seems to suggest that the BMH does not explain properly the return patterns that Song (2020) labelled as decreasing economies of scale, it appears that the BMH is present in the context of mutual funds. To confirm this interim conclusion, we perform a set of regression analyses to test the significance of the BMH by quantifying the impact that belonging to a top tercile fund portfolios has on future expected benchmark-adjusted performance estimates.

We perform three Fama Macbeth (1973) regression analyses in which we regress the arithmetic average of the next twelve months' average monthly FFC alpha estimates[10] against a dummy variable denoting whether the fund belongs to any of the top tercile fund portfolios. We use a dummy variable instead of the absolute figures of the FRR estimates to address the substantially overlapping time frames (36–47 months) of dependent (LHS) and independent (RHS) variables. In addition to the top tercile dummy variables, we include a set of control variables, namely an institutional fund dummy, average gross returns and turn ratios over the alpha estimation periods, as well as the size and the expense ratios of the funds at the start of the performance evaluation periods. We also include a size variable of the TNA at the start of the fund flow estimation period to take into account the tendency of smaller funds ending up in the top tercile fund portfolios. Finally, we include control variables for other two terciles (Middle and Bottom), as well as the proxy of displayed prior skill (FFC alpha quintiles) in the fund managers' performance. We report the results in Table IX.

The regressions show that future alpha estimates have a significant and inverse relationship with prior FRRs. Controlling for all other relevant fund characteristics, top tercile funds perform 16bps worse (1% significance level), *ceteris paribus*. The obtained coefficient estimate strengthens the earlier interim conclusion regarding the presence of the BMH, as the figure is both significant and negative, despite the earlier mentioned substantial overlap in the time frames. The coefficient estimate would be likely smaller (more negative) if we could differentiate the time frames of independent and dependent variables in a more mutually exclusive manner.

[10] Estimated over their respective prior 48 months, as per equation 1 of this thesis.

Table IX
Future factor-adjusted returns and prior factor-related returns: 1996–2019

The table presents the relationship between future benchmark-adjusted performance estimates and prior factor-related returns, elaborating the impact of luck in prior returns that are not replicated in the future. The independent variable of the regression models is the arithmetic average of FFC alphas estimated over the next twelve months using the equation 1 of this thesis. Top Tercile is a dummy variable denoting whether the fund belonged to the group of the highest FRR funds over the prior four-year period. Institutional variable is a dummy denoting if the fund share class serves institutional investors. Average gross returns are a percentage of average monthly returns over the future alpha estimation period, while expense ratio and turn ratio are annual values measured in percentages of TNA and change in portfolio composition from the start of the calendar year. $Size_t$ is the natural logarithm of TNA at the start of the future alpha estimation period, while $Size_{t-48}$ is the size of the funds at the start of the fund flow estimation period. Both size variables are earmarked TNA figures of each represented share class. The figures in parentheses are mean group standard errors for the Fama Macbeth regressions. The estimates are further controlled with dummy variables denoting Middle and Bottom Tercile funds, as well as prior four-year skill quintiles measured in FFC alpha. *, **, and *** denote statistical significance at 10%, 5%, and 1% significance level, respectively.

	(1)	(2)	(3)
Intercept (%)	-0.3242***	-0.3594***	-0.3275***
	(0.021)	(0.019)	(0.021)
Top Tercile	-0.1570***		-0.1574***
	(0.009)		(0.008)
Institutional		-0.0015	0.0022
		(0.001)	(0.001)
Average gross returns	0.4736***	-0.0070***	0.4741***
	(0.015)	(0.001)	(0.015)
Expense ratio	-0.0036*	-0.0128***	-0.0026
	(0.002)	(0.002)	(0.002)
Turn ratio	-0.0027***	-0.0070***	-0.0028***
	(0.001)	(0.001)	(0.001)
$Size_t$	-0.0120***	-0.0157***	-0.0119***
	(0.001)	(0.001)	(0.001)
$Size_{t-48}$	0.0085***	0.0114***	0.0086***
	(0.001)	(0.001)	(0.001)
Tercile controls	Yes	Yes	Yes
Skill controls	Yes	Yes	Yes
Adjusted R^2 (%)	47.96	47.67	48.02
Sample (N)	1,237,037	1,237,037	1,237,037
Periods (n)	288	288	288

Interestingly, we also find significant evidence of that the size of the funds at the start of the fund flow estimation periods impacts positively and significantly (1% significance level) the future performance estimates. Positive coefficients for the Size_{t-48} variable in all three regression models suggest, that the smaller the funds are at the start of each 48-month estimation period, the worse these funds are expected to perform during the subsequent 12-month evaluation period. This observation strengthens the hypothesis regarding that smaller funds tend to exhibit stronger BMH-associated patterns. As proposed, this is possibly a consequence of less diversified holdings. However, considering the significant coefficients for the Size_t and Turn ratio in all three models, we conclude that the evidence also supports the earlier proposed decreasing economies of scale hypothesis by Song (2020) and others.

We acknowledge, that the material overlaps in time frames between the LHS and RHS variables set limitations to the robustness of this test. However, due to the lack of better methodology, the above tests serve the purpose sufficiently. To perform a more robust analysis, we would need to have mutual fund return data on a weekly basis to have a large enough sample in estimating the future fund performance without any overlapping time frames. Since mutual funds do not report their returns more frequently than on a monthly basis, we cannot perform this test at the time of writing this thesis.

6.2 Controlling for expenses and trading activity

In addition to the BMH, the above return patterns of the Tercile portfolios can be a consequence of different trading behaviour and expense ratios of the funds. To control for the possibility of these potentially alternative explanations, we study turn ratios and expense ratios of each fund portfolio and test the significance of difference in mean between top and bottom tercile portfolios.

For the sake of comparability, we adopt similar rolling-window approach as in earlier analyses. Thus, we construct fifteen fund portfolios that we rebalance on an annual basis. For each calendar year, mutual funds are sorted into five quintiles based on their earmarked TNA at the start of the 12-month evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of prior FRRs over the preceding estimation period. We continue to compute the average TNA weighted annual expense ratios and turn ratios of the fund portfolios for the subsequent 12-month evaluation periods and

Table X
Expense and turn ratios of the fund portfolios, controlled for TNA: 1996–2019

For each calendar year, mutual funds are sorted into quintile portfolios based on the TNA at the start of the 12-month evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the same period. We then compute the average TNA weighted annual expense ratios (Panel A) and turn ratios (Panel B) for all fund portfolios. *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Parentheses report standard errors of two-way difference in means.

	Panel A: Expense ratios (%)				Panel B: Turn ratios (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1(L)	1.63	1.49	1.63	-0.01	1.14	0.94	1.04	0.10
				(0.057)				(0.067)
2	1.51	1.35	1.47	0.03	1.01	0.88	0.94	0.07
				(0.048)				(0.089)
3	1.35	1.24	1.31	0.05	0.92	0.77	0.79	0.13***
				(0.045)				(0.049)
4	1.24	1.12	1.16	0.08**	0.88	0.71	0.72	0.15**
				(0.036)				(0.075)
5(H)	0.86	0.72	0.77	0.10*	0.59	0.49	0.47	0.13**
				(0.050)				(0.055)
All	0.92	0.77	0.82	0.10**	0.64	0.51	0.51	0.13***
				(0.001)				(0.053)

compare the difference between the top and bottom tercile funds with a two-way test of difference in mean. We report the figures in Table X.

Despite that we observe some significant evidence in favour of differing ratios between the top and bottom tercile portfolios, the evidence suggests that neither the expense ratios nor the turn ratios explain the return patterns reported in Tables III–VIII. The difference between top and bottom tercile funds in expense ratios appear to be the only 10bps (5% significance level), while both top and bottom tercile funds turn only a small portion of their portfolios annually. The irrelevance of the magnitude can be illustrated with the following naïve computations: if top tercile funds have approximately 20% larger turn ratio and hence trading-related costs, considering the average trading cost of 75bps estimated by Busse, Chordia, Jiang & Tang (2017), the difference is only 15bps. Thus, the combined impact of expenses and trading patterns is approximately 25bps, being far too low compared to the underperformance figures that we reported earlier in the *Results* section of this thesis.

Surely, turn ratio is not the best proxy for the total trading volume nor trading costs of the funds, as the metric accounts for only the change in portfolio composition between the first and last trading day of each calendar year. However, assuming that mutual funds do not trade systematically front and back their entire portfolio during any given year, the metric serves as a sufficient proxy of the funds' trading behaviour over the evaluation periods.

In addition to the TNA of the funds, we also control expenses and trading costs against the factor-adjusted returns (skill) over the past 48-month. As with TNA-quintiles, the findings suggest that the earlier performance patterns are robust against the expense ratios and turn rations. We report the results in *Appendices* (Table A. IV).

6.3 Controlling for heightened idiosyncratic volatility

In addition to fund managers' inability to replicate higher lucky FRRs in the future, the earlier reported return patterns in Table III can be attributable to other idiosyncrasies of the mutual fund share classes. To control for this alternative explanation, we perform a similar analysis as in Table III but for a sub-sample period characterised with amplified levels of idiosyncratic volatility's share in return patterns of the mutual fund share class. We identify the interval of 1992–2007 as suitable for the purpose from the Figure I presented earlier in this thesis.

To maintain comparability of the results with the earlier findings and conclusions, we adopt a similar methodology in estimating and evaluating the relationship between prior FRRs and future abnormal returns. Thus, we again construct fifteen fund portfolios that we rebalance on an annual basis. Adopting the 48-month rolling window methodology, mutual funds are sorted into five quintiles based on their earmarked TNA at the start of the 12-month evaluation periods. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of prior FRRs over the 48-month estimation period. We continue to compute the average TNA weighted monthly net excess returns for the fund portfolios during the evaluation period and regress these monthly returns against the FFC model to obtain future benchmark-adjusted return estimates. We report the results in Table XI, in which the Panel A focuses on the expected net excess returns and Panel B displays benchmark-adjusted returns and OLS standard errors of the FFC alpha estimates.

Table XI
Future annualised net returns, controlled for fund TNA: 1996–2007

For each calendar year, mutual funds are sorted in two dimensions. First, mutual funds are sorted into five quintile portfolios based on their total net assets under management at the start of the 12 months long evaluation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the prior 48 months-long period. Thus, the top, middle and bottom portfolios within each quantile comprise mutual funds whose estimated factor-related average returns are among the top, middle, and bottom third of all mutual funds in that measure during the past 48 months. We then compute the average TNA weighted monthly net returns of each fund portfolio in excess of risk-free rates over the 12-month evaluation periods and annualise them. We continue by regressing these monthly average returns in excess to risk-free rate against the FFC model to obtain future benchmark-adjusted returns for each fund portfolio. Both Ret and α_{FFC} are annualised figures in percentages, while *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Figures in parenthesis are standard errors of OLS estimates.

	Panel A: Ret (%)				Panel B: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	6.17	6.25	5.84	-0.33	-2.22	-0.85	-1.21	1.01
					(1.43)	(0.82)	(1.60)	(2.46)
2	6.29	5.53	5.46	-0.83	-2.24*	-1.42**	-1.71	0.53
					(1.29)	(0.71)	(1.67)	(2.48)
3	6.67	6.62	6.91	0.25	-1.32	-0.20	0.10	1.42
					(1.52)	(0.76)	(1.63)	(2.61)
4	6.15	6.72	7.2	1.05	-2.06	-0.23	0.12	2.17
					(1.46)	(0.89)	(1.70)	(2.68)
5(H)	5.74	6.45	7.39	1.65	-1.43	0.16	1.00	2.42
					(1.14)	(0.53)	(1.75)	(2.48)
All	5.84	6.46	7.2	1.36	-1.85***	-0.51	-0.34	1.51
					(0.63)	(0.35)	(0.75)	(1.15)

Based on the above described analysis we find that the return patterns displayed in Table III are robust against the alternative explanation of fund idiosyncrasies determining the return patterns. As we see in the Table XI, mutual fund share class portfolios with the highest prior FRRs underperform in future by 185bps (1% significance level), while bottom tercile portfolios do not. The earlier significance on the fund portfolio level disappears as expected, given that the sample size falls from the full sample period considerably. Notwithstanding, the difference in between top and bottom tercile portfolios continue to remain insignificant, further piling up evidence against BMH as an alternative explanation for the decreasing economies of scale patterns that Song (2020) reports in his paper. Nevertheless, the results of the robustness analysis performed in this sub-section strengthens the legitimacy of the BMH as something future research work in mutual fund performance should control for going forward.

6.4 Controlling for the impact of the recessions

To study the impact of recessions, we remove the observations that coincide with the National Bureau of Economic Research (NBER) recessions and perform similar analyses as in Tables II–VIII. Performing these analyses, we find that the reported patterns in Table III are in fact materially driven by the recessions. Analysing the performance difference between top and bottom tercile funds, we find that any potential performance difference between them vanishes once we remove the recession periods from the sample. Thus, it appears that the consistency of the inverse relationship between prior FRRs and future abnormal returns is not robust over time.

However, given that the recessions are part of the probability distribution that drives the data generation process of mutual funds returns, we find the implications of this robustness test to be elaborating rather than dispositioning by nature. For instance, it seems that mutual funds' fire sales and hence liquidity demands could explain some portion of the return patterns that we observe top and bottom terciles. The Tables of the above analyses are reported in *Appendices* (Tables A. V–VI).

7 Conclusion

In this thesis, we study the role of the mutual fund investors' inability to recognise fund managers' limited abilities to replicate prior lucky returns in the future as a source of the decreasing economies of scale patterns that Song (2020) reports in his paper. Empirically, we find that although mutual fund share classes appear to attract excess fund flows measured in absolute dollar value, we do not find evidence of significant excess fund flows when we control the flows against the size of the mutual fund share classes at the start of the respective evaluation periods. Additionally, we find that mutual fund share classes associated with high realised FRRs do not outgrow other share classes. If anything, we find that share classes with the highest FRRs often represent the lower end of the spectrum of share classes measured in earmarked total net assets at the start of the estimation periods, implying that sub-optimal diversification and increased concentration in risk exposures might orchestrate which share classes transpire among the top third of estimated FRRs in a given time period.

Despite the above insights related to size and fund flows, interestingly, we find that mutual funds with the highest prior FRRs seem underperform significantly their respective benchmarks in the future while share classes with the lowest prior FRRs do not. Moreover, the sophistication that prevails in the funds' investor bases seems to impact to which extent the share classes exhibit BMH-associated return patterns. Regardless of whether the fund share classes serve retail or institutional funds, the underperformance of the share classes with the highest prior FRRs is significant and substantial, although institutional fund share classes exhibit BMH-associated return patterns that are only half of those compared to the figures of the corresponding retail fund share classes. The findings suggest that investors' lack of skill in benchmarking is linked with the phenomenon.

However, despite the above favourable patterns from the perspective of the BMH, we find consistently insignificant evidence for a performance difference between fund share classes that are associated with the highest and the lowest prior FRRs. Furthermore, even if the performance difference were statistically significant, the estimated performance difference of 158bps is not high enough to explain the significant difference of 300–400bps that Song (2020) reports in his paper studying the phenomenon by aggregating separate share classes into single mutual funds and orthogonalising the risk factors in his benchmarking models.

The contribution of this thesis is threefold. First, we show that despite favourable patterns from the perspective of the BMH, we do not find statistically powerful enough evidence to disposition earlier conclusions that propose the reported future risk-adjusted underperformance to be a consequence of increasing liquidity demands of the funds. In fact, we even find occasionally evidence that aligns with the decreasing economies of scale hypothesis. Hence, the findings of this thesis increases the robustness of Song's (2020) earlier conclusions. Secondly, since we display consistency in patterns that align with the BMH, we emphasise implicitly that vantage point of mutual fund research going forward. Although the phenomenon does not appear to exhibit conclusive role in mutual funds' decreasing economies of scale hypothesis, the findings nevertheless suggest that future fund performance studies should control for the possibility of luck in preceding realised return patterns more granularly. Third, we find new contradicting evidence with the earlier findings regarding the mutual fund flow patterns. We find that when we analyse the size-adjusted fund flows on share class level, we find no evidence of irrational excess fund flows that are attributable to higher realised FRRs of the fund share classes. Thus, it seems that the fund flow dynamics are more rational compared to what has been concluded as of late on complete mutual fund level. Overall, the findings of this thesis are relevant for many mutual fund industry stakeholders, including academia, non-professional mutual fund retail investors, and institutional asset managers.

7.1 Limitations and further research potential

Some limitations diminish the value of the results. The majority of the limitations are related to not controlling for phenomena that Song (2020) controlled for in his paper. First, we have not taken into account a phenomenon documented by Kamstra, Kramer, Levi & Wermers (2017). They find that flows into equity mutual funds from other fund categories, such as fixed-income funds and money market funds, occur more likely in spring. Second, we do not test the robustness of the results either against the CRSP-classified cap-based funds or style-based funds.

In addition to those, some distributional parameters of the sample can introduce limitations into the findings. For instance, performing χ^2-test of normality we find significant evidence against the normality of the monthly returns, FFC alpha estimates, and estimated quarterly fund flows. The evidence against the normality of these sample distributions can introduce biases in the study.

Finally, we do not control for several other theories that have been documented empirically to prevail in the mutual fund industry. These theories include *slow moving capital*

(Duffie, 2010) and *closet indexing* (Petäjistö, 2013). Therefore, interesting further research avenues could be determining to which extent these phenomena relate to the reported patterns in this thesis as well as the findings proposed by Song (2020).

References

Barber, B. Huang, X., Odean, T. 2016. Which factors matter to investors? Evidence from mutual fund flows. Review of Financial Studies. Vol 29. P. 2600–2642.

Ben-David, I. Li, J. Rossi, A. Song, Y. 2018. What Do Mutual Fund Investors Really Care About? Working paper.

Berk, J. Green, R. 2004. Mutual fund flows and performance in rational markets. Journal of Political Economy. Vol 112. P. 1269–1295.

Berk, J., van Binsbergen, J. 2016. Assessing asset pricing models using revealed preference. Journal of Financial Economics. Vol 119. P. 1–23.

Bollen, N. Joenväärä, J. Kauppila, M. 2020. Hedge Fund Performance: End of an Era? Working paper.

Busse, J. Chordia, T. Jiang, L. Tang, Y. 2017. Mutual Fund Trading Costs and Diseconomies of Scale. Working paper. Singapore Management University.

Carhart, M. 1997. On persistence in mutual fund performance. Journal of Finance. Vol 52. P. 57– 82.

Chakraborty, I. Kumar, A. Muhlhofer, T. Sastry, R. 2018. Does limited investor attention explain mutual fund flows? Evidence from sector funds. Working paper, University of Miami.

Chen, J. Harrison, H. Huang, M. Kubik, J. 2004. Does fund size erode mutual fund performance? The role of liquidity and organisation. American Economic Review. Vol 94. P. 2941–2969.

Cooper, M. Gulen, H. Rau, P. 2005. Changing Names with Style: Mutual Fund Name Changes and Their Effects on Fund Flows. The Journal of Finance. Vol 60:6. P. 2825–2858.

Del Guercio, D. Tkac, P. 2008. Star power: The effect of Morningstar ratings on mutual fund flows. Journal of Financial and Quantitative Analysis. Vol 43:3. P. 907–936.

Duffie, D. 2010. Presidential address: Asset price dynamics with slow-moving capital. Journal of Finance. Vol 65:4. P. 1237–1267.

Edelen, R. Evans, R. Kadlec, G. 2009. Scale effects in mutual fund performance: The role of trading costs. Working paper, University of Virginia.

Edelen, R. Evans, R. Kadlec, G. 2013. Shedding light on “invisible” costs: Trading costs and mutual fund performance. Financial Analysts Journal. Vol 69. P. 33–44.

Fama, E. French, K. 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics. Vol 33. P. 3–56.

Friesen, G. Sapp, T. 2007. Mutual fund flows and investor returns: An empirical examination of fund investor timing ability. Journal of Banking and Finance. Vol 31:9. P. 2796–2816.

Grinblatt, M. Keloharju, M. Linnainmaa, J. 2012. IQ, trading behaviour, and performance. Journal of Financial Economics. Vol 104:2. P. 339–362.

Grinblatt, M. Titman, S. 1989. Mutual fund performance: An analysis of quarterly portfolio holdings. Journal of Business. Vol 62. P. 393–416.

Harvey, C., Liu, Y. 2017. Decreasing returns to scale, fund flows, and performance. Working paper, Duke University.

Investment Company Institute. 2019. Investment Company Factbook – A Review of Trends and Activities in the Investment Company Industry. 60th edition.

Jain, P. Wu, J. 2000. Truth in Mutual Fund Advertising: Evidence on Future Performance and Fund Flows. The Journal of Finance. Vol 55:2. P. 937–958.

Kamstra, M., Kramer, L. Levi, M. Wermers, R. 2017. Seasonal asset allocation: Evidence from mutual fund flows. Journal of Financial and Quantitative Analysis. Vol 52. P. 71– 109.

Lou, D. 2012. A flow-based explanation for return predictability. Review of Financial Studies. Vol 25. P. 3457–3489.

Muñoz, F. Vicente, R. 2018. Hindsight effect: What are the actual cash flow timing skills of mutual fund investors? Journal of Empirical Finance. Vol 45. P. 181–193.

Pástor, L. Stambaugh, R. 2002a. Investing in equity mutual funds. Journal of Financial Economics. Vol 63. P. 351–380.

Pástor, L. Stambaugh, R. Taylor, L. 2020. Fund trade-offs. Journal of Financial Economics. In Press.

Petäjistö, S. 2013. Active share and mutual fund performance. Financial Analyst Journal. Vol 69:4. P. 73–93.

Pollet, J., Wilson, M. 2008. How does size affect mutual fund behaviour? Journal of Finance. Vol 63. P. 2941– 2969.

Pronin, N. 2020. Think you can choose your asset manager? Think again – mutual fund investors' inability to distinguish fund managers' skill. Bachelor's thesis. Aalto University School of Business, Faculty of Finance.

Solomon, D., Soltes, E. Sosyura, D. 2014. Winners in the spotlight: Media coverage of fund holdings as a driver of flows. Journal of Financial Economics. Vol 113. P. 53–72.

Siegel, L.B. Waring, M.B. and Scanlan, M.H. 2009. Five principles to hold onto (even when your boss says the opposite). The Journal of Portfolio Management, 35(2). P.25–41.

Sirri, E. Tufano, P. 1998. Costly Search and Mutual Fund Flows. The Journal of Finance. Vol 53:5. P. 1589–1622.

Yan, X. 2008. Liquidity, investment style, and the relation between fund size and fund performance. Journal of Financial and Quantitative Analysis. Vol 43. P. 741–768.

Song, Y. 2020. The Mismatch Between Mutual Fund Size and Skill. Journal of Finance. Vol LXXV, No 5. P. 2555–2589.

Zhu, M. 2018. Informative fund size, managerial skill, and investor rationality. Journal of Financial Economics. Vol 130. P. 114–134.

Appendices

Figure A.I: FED balance sheet and appreciation of equity indices

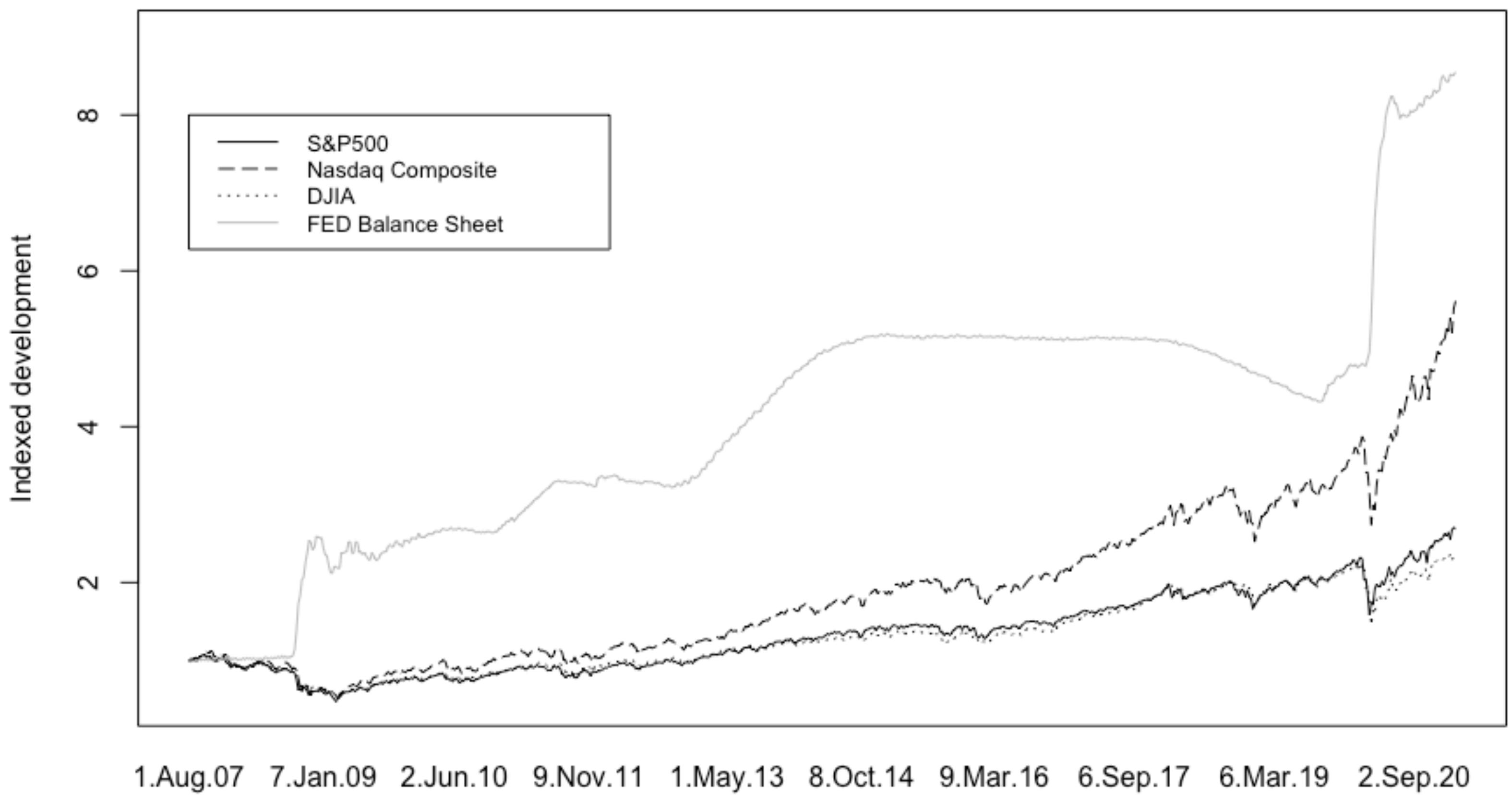


Table A. I
Future annualised net returns, controlled for skill: 1996–2019

For each calendar year, mutual funds are sorted in two dimensions. First, each mutual fund is sorted into five quintile portfolios based on their factor-adjusted returns (FFC alpha) during the 48 months long estimation period. Each quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the same 48 months long period. Thus, the top, middle and bottom tercile portfolios within each quantile comprise mutual funds whose estimated factor-related average returns during the past 48 months are among the top, middle, and bottom third of all mutual funds in that measure. We then compute the average TNA weighted monthly net returns of each fund portfolio in excess of risk-free rates over the 12-month evaluation periods and annualise them. We continue by regressing the monthly average returns in excess to risk-free rate against the FFC model to obtain future benchmark-adjusted expected returns for the fund portfolios. Both Ret and α_{FFC} are annualised figures in percentages, while *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Figures in parentheses are standard errors of OLS estimates.

	Panel A: Ret (%)				Panel B: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	6.86	6.64	5.87	-0.99	-1.71*	-1.21*	-2.17**	-0.46
					(0.93)	(0.64)	(1.01)	(1.33)
2	6.36	8.05	6.86	0.50	-1.92**	0.47	-0.97	0.95
					(0.81)	(1.06)	(0.97)	(1.48)
3	6.85	6.64	7.38	0.53	-1.11	-0.60	-0.31	0.81
					(0.78)	(0.44)	(0.94)	(1.34)
4	7.59	7.22	8.25	0.66	-0.34	-0.10	0.60	0.93
					(0.79)	(0.44)	(1.00)	(1.45)
5(H)	6.08	6.36	7.64	1.56	-2.06**	-0.63	0.43	2.49
					(1.00)	(0.65)	(1.13)	(1.65)
All	6.93	6.99	7.50	0.56	-1.43***	-0.41	-0.48	0.94
					(0.40)	(0.32)	(0.47)	(0.69)

Table A. II
The response of mutual fund flows to factor-related returns: 1992–2007

For each calendar year of data, all mutual funds are sorted into five quintile portfolios based on their four-factor alpha in prior 48 months. Each alpha-portfolio is then divided into three tercile portfolios based on their order in sample distribution of factor-related returns during the full sample period. Thus, the top, middle, and bottom tercile portfolios represent a portfolio of mutual funds whose estimated factor-related average past 48-month returns are in the top, middle, and bottom third of all mutual funds. A_s and A_e (millions in U.S. dollars) are the time series averages of cross-sectional arithmetic means of total net assets under management for a given fund portfolio at the start and the end of each 48-month estimation period. Flow (in millions) is the time series averages of the cross-sectional mean of average quarterly flows over each same 48-month estimation period. Δ and α^{CAPM} (in %) are the annualised time series averages of the cross-sectional mean of the factor-related returns and CAPM alphas, respectively. t-stat is the t-statistic of a test of difference in average quarterly flows between the top and bottom tercile portfolios within a given quintile. *, **, and *** denote 10%, 5%, and 1% statistical significance, respectively.

	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	t -stat
	Top Tercile Δ					Middle Tercile Δ					Bottom Tercile Δ					Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1 (L)	254	312	-2.4	6.5%	-2.1%	343	262	-4.0	1.7%	-5.7%	462	245	-5.3	-3.0%	-10.9%	4.35***
2	279	521	3.1	5.9%	2.50%	384	421	-2.5	1.5%	-1.4%	672	469	-5.8	-2.5%	-5.2%	10.2***
3	401	915	6.5	5.7%	4.10%	527	748	1.9	1.4%	0.3%	647	632	0.1	-2.3%	-3.0%	6.31***
4	325	990	10.6	5.9%	6.10%	819	1,420	8.1	1.5%	2.2%	1,014	1,229	5.0	-2.4%	-1.6%	5.09***
5 (H)	210	822	13.4	7.2%	14.7%	428	1,114	15.2	1.7%	6.5%	410	796	12.4	-3.6%	2.7%	0.88

Table A. III
The response of mutual fund flows to factor-related returns: 2010–2019

For each calendar year of data, all mutual funds are sorted into five quintile portfolios based on their four-factor alpha during prior 48 months. Each alpha-portfolio is then divided into three tercile portfolios based on their order in sample distribution of factor-related returns during the full sample period. Thus, the top, middle, and bottom tercile portfolios represent a portfolio of mutual funds whose estimated factor-related average past 48-month returns are in the top, middle, and bottom third of all mutual funds. A_s and A_e (millions in U.S. dollars) are the time series averages of cross-sectional arithmetic means of total net assets under management for a given fund portfolio at the start and the end of each 48-month estimation period. Flow (in millions) is the time series averages of the cross-sectional mean of average quarterly flows over each 48-month estimation period. Δ and α^{CAPM} (in %) are the annualised time series averages of the cross-sectional mean of the factor-related returns and CAPM alphas, respectively. t-stat is the test statistic of a test of difference in average quarterly flows between the top and bottom tercile group within a given quintile. *, **, and *** denote 10%, 5%, and 1% statistical significance, respectively.

	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	t -stat
	Top Tercile Δ					Middle Tercile Δ					Bottom Tercile Δ					Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1 (L)	319	223	-6.2	2.0%	-4.0%	357	262	-6.9	0.3%	-5.1%	330	208	-6.3	-1.4%	-8.5%	0.22
2	390	362	-6.3	1.6%	-1.3%	382	400	-3.6	0.2%	-2.3%	436	394	-6.7	-0.8%	-3.7%	0.79
3	462	543	-4.5	1.4%	-0.3%	532	650	-1.1	0.3%	-1.1%	778	849	-3.2	-0.8%	-2.4%	-1.96*
4	725	1,077	-2.4	1.4%	0.9%	873	1,222	-0.4	0.3%	0.1%	700	951	-1.5	-0.9%	-1.4%	-1.17
5 (H)	568	915	2.5	1.7%	3.7%	740	1,202	5.3	0.4%	2.6%	632	909	2.6	-1.2%	1.5%	-0.16

Table A. IV

Expense and turn ratios of the fund portfolios, controlled for skill: 1996–2019

For each calendar year, mutual funds are sorted into quintile portfolios based on their factor-adjusted returns (FFC alpha) over the past 48 months estimation period. Each skill quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the same period. We then compute the average TNA weighted annual expense ratios (Panel A) and turn ratios (Panel B) for all fund portfolios. *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Parentheses report standard errors of two-way difference in means.

	Panel A: Expense ratios (%)				Panel B: Turn ratios (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1(L)	1.14	1.13	1.11	0.27***	0.91	0.86	0.83	0.09
				(0.046)				(0.086)
2	1.05	0.94	0.98	0.07	0.78	0.78	0.63	0.14**
				(0.047)				(0.067)
3	0.97	0.76	0.79	0.17***	0.70	0.46	0.49	0.21***
				(0.057)				(0.083)
4	0.80	0.66	0.69	0.11	0.54	0.37	0.39	0.15**
				(0.072)				(0.066)
5(H)	0.98	0.86	0.94	0.03	0.59	0.64	0.60	-0.01
				(0.053)				(0.057)
All	0.92	0.77	0.82	0.10**	0.64	0.51	0.51	0.13***
				(0.051)				(0.053)

Table A. V
Fund flows and factor-related returns, excluding recessions: 1992–2019

For each calendar year of data, all mutual funds are sorted into five quintile portfolios based on their four-factor alpha in prior 48 months. Each alpha-portfolio is then divided into three tercile portfolios based on their order in sample distribution of factor-related returns during the full sample period. Thus, the top, middle, and bottom tercile portfolios represent a portfolio of mutual funds whose estimated factor-related average past 48-month returns are in the top, middle, and bottom third of all mutual funds. A_s and A_e (millions in U.S. dollars) are the time series averages of cross-sectional arithmetic means of total net assets under management for a given fund portfolio at the start and the end of each 48-month estimation period. Flow (in millions) is the time series averages of the cross-sectional mean of average quarterly flows over 48-month estimation periods. Δ and α^{CAPM} (in %) are the annualised time series averages of the cross-sectional mean of the factor-related returns and CAPM alphas, respectively. t-stat is the t-statistic of a test of difference in average quarterly flows between the top and bottom tercile portfolios within a given quintile. *, **, and *** denote 10%, 5%, and 1% statistical significance, respectively.

	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	A_s	A_e	Flow	Δ	α^{CAPM}	t -stat
	Top Tercile Δ					Middle Tercile Δ					Bottom Tercile Δ					Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1 (L)	306	260	-4.9	3.6%	-3.1%	362	266	-6.0	0.7%	-5.3%	370	219	-6.0	-1.7%	-9.1%	2.79***
2	357	408	-3.5	2.9%	0.1%	386	401	-3.3	0.6%	-2.1%	511	413	-6.2	-1.3%	-4.2%	5.18***
3	447	629	-1.9	2.5%	0.9%	540	672	-0.5	0.6%	-0.7%	754	782	-1.8	-1.3%	-2.7%	-0.20
4	629	1,055	0.6	2.6%	2.3%	870	1,255	1.7	0.7%	0.7%	827	1,037	0.7	-1.3%	-1.5%	-0.17
5 (H)	467	877	5.4	3.4%	6.8%	665	1,179	8.1	0.8%	3.8%	578	871	5.5	-1.9%	1.8%	-0.22

Table A. VI
Future annualised net returns, excluding recessions: 1996–2019

For each calendar year, mutual funds are sorted in two dimensions. First, each mutual fund is sorted into five quintile portfolios based on their total net assets under management at the start of the 12 months long observation period. Each TNA quintile portfolio is then divided into three tercile portfolios based on the sample distribution of factor-related average returns during the prior 48 months-long period. Thus, the top, middle and bottom portfolios within each quantile comprise mutual funds whose estimated factor-related average returns during the past 48 months are among the top, middle, and bottom third of all mutual funds in that measure. We then compute the average TNA weighted monthly net returns of each fund portfolio in excess of risk-free rates over the 12-month evaluation periods and annualise them. We continue by regressing these monthly average returns in excess to risk-free rate against the FFC model to obtain future benchmark-adjusted returns for each fund portfolio. Both Ret and α_{FFC} are annualised figures in percentages, while *, **, and *** denote statistical significance at 10%, 5%, and 1% level, respectively. Figures in parentheses are standard errors of OLS estimates.

	Panel A: Ret (%)				Panel B: α_{FFC} (%)			
	Top	Middle	Bottom	Diff	Top	Middle	Bottom	Diff
1(L)	9.62	8.80	8.18	-1.44	-1.74*	-1.24**	-1.96*	-0.22
					(0.98)	(0.59)	(1.01)	(1.61)
2	9.85	8.58	8.11	-1.74	-1.49	-1.39**	-1.93*	-0.44
					(0.91)	(0.56)	(1.07)	(1.63)
3	10.09	9.44	9.02	-1.07	-1.22	-0.43	-1.14	0.08
					(0.92)	(0.57)	(1.03)	(1.66)
4	10.17	9.65	9.41	-0.76	-1.39	-0.50	-0.99	0.40
					(0.97)	(0.64)	(1.09)	(1.74)
5(H)	10.35	9.88	10.12	-0.23	-0.98	-0.10	-0.20	0.78
					(0.73)	(0.45)	(1.13)	(1.62)
All	10.34	9.85	9.90	-0.44	-1.37***	-0.73***	-1.24**	0.12
					(0.41)	(0.26)	(0.48)	(0.75)